UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Estimação da creatinina sérica basal através de modelos GAMLSS

Inês Rodrigues Mendonça

Mestrado em Bioestatística

Dissertação orientada por: Prof.ª Doutora Maria Fernanda Nunes Diamantino e Prof.ª Doutora Ana Luísa Trigoso Papoila da Silva

2016

Agradecimentos

- ◇ À Professora Doutora Ana Luísa Papoila pelo apoio, orientação, disponibilidade e ensinamentos que me transmitiu ao longo deste trabalho. Agradeço-lhe imenso todo o seu esforço para garantir o meu ótimo trabalho.
- ◇ À Professora Doutora Fernanda Diamantino pela orientação, apoio, motivação e ensinamentos, não só durante este trabalho, mas ao longo do mestrado.
- ◇ À Doutora Karina Soto pela disponibilidade dos dados utilizados neste trabalho.
- ◇ À Professora Doutora Lisete de Sousa pelo apoio e incentivo durante o mestrado.
- ◇ Aos meus pais e irmão, porque sem eles nunca teria chegado onde cheguei. Quero agradecer-lhes todo o apoio que me deram ao longo da vida, em especial durante o meu percurso académico. Obrigada por não me deixarem desistir ao primeiro sinal de dificuldade. Espero que tenham orgulho em mim como eu tenho de vocês.

Resumo

A doença renal aguda é caracterizada pela rápida diminuição da função renal. Esta patologia, por vezes, pode ser assintomática em alguns doentes, apresentando apenas variações de parâmetros laboratoriais que avaliam a taxa de filtração glomerular.

O diagnóstico da lesão renal aguda é feito através de biomarcadores renais como a creatinina sérica. Este valor, juntamente com critérios de classificação da lesão renal aguda, consegue determinar a severidade da doença. No entanto, a classificação baseia-se em alterações da creatinina sérica em relação ao valor da creatinina sérica basal que frequentemente não está disponível. Assim, a abordagem utilizada na área clínica passa pela estimação dos valores da creatinina sérica basal através de várias fórmulas para o efeito.

Neste estudo serão abordados três métodos de regressão, na tentativa de encontrar um modelo com bom poder preditivo que estime o valor da creatinina sérica basal. Assim sendo, o objetivo deste estudo passa pela aplicação de modelos lineares generalizados, modelos aditivos generalizados e modelos GAMLSS. Os modelos são avaliados pela sua capacidade preditiva através da análise dos resíduos.

Uma vez que, a distribuição dos resíduos obtidos pelos modelos lineares generalizados e pelos modelos aditivos generalizados não foi normal, houve necessidade de prosseguir para uma nova abordagem baseada nos GAMLSS. Desta forma, foi alcançado o pressuposto de normalidade dos resíduos obtidos por estes modelos, embora as estimativas obtidas para os valores da creatinina sérica basal não tenham sido as melhores.

Palavras-Chave: doença renal aguda, creatinina sérica basal, modelos lineares generalizados, modelos aditivos generalizados, GAMLSS, resíduos.

Abstract

Acute renal disease is characterized by a rapid decrease of the renal function. This condition can, for some patients, be asymptomatic only changing for some laboratory parameters such as the glomerular filtration rate.

The diagnosis of acute kidney injury is made by renal biomarkers such as serum creatinine. This value, along with the acute kidney injury classification criteria, can define the severity of the disease. However, the classification is based on changes in serum creatinine, when compared with a baseline value, which, most of the times, is not available. These values are replaced by estimates obtained through appropriate formulae.

This study will present three regression approaches in an attempt to find a model with a good predictive power that is able to estimate the value of the baseline serum creatinine. Therefore, the goal of this work involves the application of generalized linear models, generalized additive models and GAMLSS models. These models are evaluated by their predictive ability, through a residuals analysis.

The residuals of the generalized linear models and generalized additive models violated the assumption of normality and further analysis was needed, using the GAMLSS, to obtain a predictive model for the estimation of the baseline serum creatinine. With these models the normality of the residuals was achieved although the obtained estimates for baseline serum creatinine were not the best.

Keywords: acute renal disease, serum creatinine baseline, generalized linear models, generalized additive models, GAMLSS, residuals.

Índice

Lista de Tabelas

Lista de Figuras

Siglas e Acrónimos

ADQI -Acute Dialysis Quality Initiative
AIC - Akaike Information Criteria
AKI - Acute Kidney Injury (lesão renal aguda)
AKIN - Acute Kidney Injury Network
CG - generalização do algoritmo de Cole e Green
CKD - Chronic Kidney Disease (doença renal crónica)
cs - cubic smoothing splines
CV - Cross Validation
d.f. - degrees of freedom (número de graus de liberdade)
f.d.p. - função densidade de probabilidade
GAIC - Generalized Akaike Information Criteria
GAMLSS - Generalised Additive Models for Location, Scale and Shape
GB1 - Generalized Beta type 1
GB2 - Generalized Beta type 2
GCV - Generalised Cross Validation
GD - Global Deviance
GFR - Glomerular Filtration Rate (taxa de filtração glomerular)
HFF - Hospital Professor Doutor Fernando Fonseca
KDIGO - Kidney Disease Improving Global Outcomes
MAG - Modelos Aditivos Generalizados
MDRD - Modification of Diet in Renal Disease
MLG - Modelos Lineares Generalizados
NHANES III - Third National Health and Nutrition Examination Survey
P-splines - Penalized smoothing splines
PWLS - Penalised Weighted Least Squares
RIFLE - Risk, Injury, Failure, Loss-of-function and End-stage kidney disease
RS - generalização do algoritmo de Rigby e Stasinopoulos
SCr - Serum Creatinine (creatinina sérica)
WLS - Weighted Least Squares

Capítulo 1

Introdução

1.1 O Rim

Uma das principais funções do rim é a remoção da corrente sanguínea de produtos metabólicos desnecessários e de substâncias ingeridas em excesso. Produtos nitrogenados como por exemplo, a ureia ou a creatinina, produzidos pelo metabolismo, são removidos do organismo através dos rins. A taxa média de excreção dos produtos metabólicos contabiliza as concentrações dos produtos removidos da corrente sanguínea, com o objetivo de manter os seus equilíbrios no organismo (Eaton and Pooler, 2009).

A *clearance* renal significa a remoção de substâncias da corrente sanguínea através de uma série de mecanismos renais. A taxa a que uma determinada substância é excretada do corpo num certo período de tempo ou o tempo que a sua concentração demora até ficar reduzida a metade são habitualmente utilizadas para quantificar a *clearance* (Eaton and Pooler, 2009).

O volume de filtração plasmática (plasma filtrado pelo rim), em função de uma unidade de tempo, é quantificado pela taxa de filtração glomerular (GFR − *Glomerular Filtration Rate*). Esta taxa é crucial para determinar a função renal que nem sempre é constante, variando consoante diversos estados fisiológicos e certas patologias (Eaton and Pooler, 2009).

O *goldstandard* para avaliar a GFR é através da *clearance* da *inulin* (polissacarídeo pertencente ao grupo dos hidratos de carbono), que é uma substância utilizada em diversos estudos de investigação ou situações clínicas, quando é necessário um valor preciso da GFR. No entanto, este método de avaliação da *inulin* é complicado e, por isso, na prática clínica comum é avaliada a creatinina sérica (SCr) como indicador da GFR (Eaton and Pooler, 2009).

Em circunstâncias normais, a concentração de creatinina sérica mantem-se praticamente constante devido ao equilíbrio entre a sua produção e a sua excreção. No entanto, um aumento do seu valor plasmático é sinal de que pode haver algum problema a nível renal (Eaton and Pooler, 2009).

A concentração da ureia plasmática também pode ser utilizada para avaliar a GFR. No entanto, é um biomarcador menos preciso que a creatinina sérica, uma vez que durante o seu processo de excreção existe uma fase de reabsorção. Além disso a ureia sérica varia consoante a

ingestão proteica, o catabolismo tecidual e o controlo da regulação hormonal (Eaton and Pooler, 2009).

1.2 Creatinina Sérica

A creatinina sérica deriva, principalmente, do metabolismo da creatinina muscular, e está relacionada proporcionalmente com a massa muscular do indivíduo. Esta é diferente entre os sexos e as idades, levando a diferentes concentrações de produção de creatinina sérica. Valores mais elevados de creatinina sérica são mais comuns em indivíduos do sexo masculino do que em indivíduos do sexo feminino. No caso dos indivíduos jovens, os valores de creatinina sérica são normalmente maiores do que os valores dos indivíduos idosos. Os valores de creatinina sérica também são influenciados pela raça. No caso dos indivíduos de raça negra, pelo facto de possuírem mais massa muscular do que os de raça não negra, têm valores superiores de creatinina sérica (Abensur, 2011).

Existem diversos fatores que influenciam a concentração da SCr: percentagem de massa muscular, consumo de carne vermelha, lesão muscular (exemplo: rabdomiólise − deterioração rápida do músculo esquelético), diminuição da secreção tubular, entre outras (Kumar and Clark, 2009). Outros fatores como a idade, sexo, metabolismo de fármacos, dieta ou exercício físico também influenciam as concentrações da creatinina sérica (Solomon and Segal, 2008). Um outro fator importante que influencia a concentração da SCr é a raça. De facto, através de um estudo realizado nos EUA, pela *Third National Health and Nutrition Examination Survey (NHANES III)*, concluiu-se que indivíduos de raça negra têm um maior valor de creatinina sérica basal comparativamente a indivíduos de raça não negra, independentemente da sua função renal, sexo ou idade (Hsu et al., 2008).

O intervalo de valores admissíveis para a creatinina sérica, considerados normais pela maioria dos laboratórios de análises clínicas, é entre 0.6 - 1.3 mg/dL (Abensur, 2011).

A creatinina sérica é um bom parâmetro clínico de avaliação da GFR porque: (1) a excreção de creatinina tem uma boa correlação com a determinação da GFR pela *inulin*; (2) a excreção da creatinina é relativamente constante durante o dia; (3) a determinação da creatinina sérica é um procedimento simples e realizado na maioria dos laboratórios de análises clínicas (Abensur, 2011).

1.3 Doença Renal Aguda

A lesão renal aguda (*Acute Kidney Injury* - AKI), previamente conhecida como falha renal aguda, (*acute renal failure*) é caracterizada pela rápida diminuição da função renal, resultando na acumulação de compostos azotados (nitrogénio metabolizado, como ureia e creatinina) e resíduos que normalmente são eliminados pelos rins. A lesão renal aguda não é apenas uma só patologia, mas sim um grupo heterogéneo de condições que partilham certas condições de diagnóstico: um aumento da concentração sérica de ureia e/ou um aumento da concentração sérica da creatinina, que normalmente está associado à diminuição do volume de urina produzida

(Longo et al., 2012).

A AKI pode manifestar-se através de um grande número de patologias com elevada severidade. No entanto, alguns casos podem ser assintomáticos com apenas alterações passageiras nos parâmetros laboratoriais que determinam a GFR. Estas alterações são, por vezes, significativas, e podem tornar-se rapidamente fatais (Longo et al., 2012).

A incidência da AKI tem vindo a aumentar a nível global (Soto et al., 2010). A AKI afeta cerca de 10 a 20%, dos adultos hospitalizados (Levey et al., 2015).

Esta patologia é uma causa comum de mortalidade, morbilidade e um importante fator de risco de progressão de doenças renais para fase terminal, como a doença renal crónica (CKD − *Chronic Kidney Disease*) que afeta cerca de 10% de adultos não hospitalizados (Levey et al., 2015). A progressão destas patologias leva a uma falha renal que pode ter que recorrer a tratamento por diálise ou transplante renal (Siew et al., 2010).

Detetar a AKI em estágios iniciais da doença facilita o seu diagnóstico e tratamento. No entanto, no início desta patologia podem não ser detetados sinais ou sintomas no doente. Como alternativa, recorre-se a valores laboratoriais para efetuar o diagnóstico da AKI (Levey et al., 2015).

1.4 Diagnóstico da Doença Renal Aguda

As técnicas de diagnóstico da AKI utilizam biomarcadores renais como a creatinina (Waikar and Bonventre, 2009). Este biomarcador é o mais utilizado na prática clínica sendo considerado o *gold standart*, juntamente com o *output* (débito) urinário, para o diagnóstico da AKI. Uma vez que a creatinina sérica não é um biomarcador de lesão, não basta apenas ser detetado na corrente sanguínea para confirmar o diagnóstico da AKI (Waikar et al., 2009).

A classificação da AKI baseia-se nas alterações da concentração da creatinina sérica, em relação a um valor basal, e na análise do *output* urinário (Závada et al., 2010). Os critérios RIFLE (*Risk, Injury, Failure, Loss-of-function and End-stage kidney disease*) e AKIN (*Acute Kidney Injury Network*) tentaram normalizar o diagnóstico e a classificação da AKI pela utilização das variações deste biomarcador e dos diferentes graus de progressão da oligúria (diminuição da produção de urina), visualizado no *output* urinário (Figura 1.1) (Siew et al., 2010).

O critério RIFLE, (Figura 1.1), introduzido em 2004 pelo *Acute Dialyses Quality Initiativa Working Group* - (ADQI), divide a AKI em três estágios de severidade - Risco, Lesão e Falha - e em dois *outcomes* clínicos - Perda renal e Estágio-final da doença renal. Este critério baseia-se em alterações relativas da creatinina sérica, na taxa de filtração glomerular e no *output* urinário (Bagshaw et al., 2009). Por exemplo, se ocorre um aumento de duas vezes o valor da creatinina sérica basal e uma diminuição de $< 0.5ml/kg$ por hora, durante 12 horas do *output* urinário, o doente é classificado no estágio de 'Lesão' (Leung et al., 2013).

Em 2007, foi introduzido um novo critério pelo *Acute Kidney Injury Network* (AKIN), (Figura 1.1), que propôs modificações do critério RIFLE, incluindo um *threshold* de variações absolutas dos valores de creatinina sérica (Bagshaw et al., 2009). Por exemplo, um aumento da creatinina sérica $\geq$ 0.3 mg/dL, em relação à creatinina sérica basal, juntamente com a diminuição do *output* urinário $< 0.5ml/kg$ por hora, num período inferior a 6 horas, classifica a AKI no estágio 1 da patologia (Leung et al., 2013).

Mais recentemente, em 2013, foi introduzido o critério KDIGO - *Kidney Disease Improving Global Outcomes*, (Figura 1.1), que também utiliza os valores da creatinina sérica, em relação a um valor basal, para a classificação da AKI. Este critério de classificação, como AKIN, também divide o estágio de gravidade da AKI em 3 estágios de severidade. Para além disso também utiliza o *output* urinário para ajudar na classificação (Leung et al., 2013).

Table 1 | AKI criteria

Classification*	Serum creatinine, GFR or other criteria	Urine output criteria
***KDIGO*[67]**		
Stage 1	1.5–1.9-fold increase in serum creatinine level from baseline or ≥0.3 mg/dl (26.5 μmol/l) increase in serum creatinine level	<0.5 ml/kg per h for 6–12 h
Stage 2	2.0–2.9-fold increase in serum creatinine level from baseline	<0.5 ml/kg per h for ≥12 h
Stage 3	Threefold increase in serum creatinine level from baseline or increase in serum creatinine level to ≥4.0 mg/dl (≥353.6 μmol/l) or initiation of renal replacement therapy or, in patients aged <18 years, decrease in eGFR to <35 ml/min/1.73 m²	<0.3 mg/kg per h for ≥24 h or anuria for ≥12 h
***AKIN*[15]**		
Stage 1	Increase in serum creatinine level of ≥0.3 mg/dl (≥26.4 μmol/l) or increase to ≥150–200% (1.5–2.0-fold) from baseline	<0.5 ml/kg per h for >6 h
Stage 2	Increase in serum creatinine level to >200–300% (>2–3-fold) from baseline	<0.5 ml/kg per h for >12 h
Stage 3	Increase in serum creatinine level to >300% (>3-fold) from baseline (or serum creatinine level of ≥4.0 mg/dl [≥354 μmol/l] with an acute increase of at least 0.5 mg/dl [44 μmol/l])	<0.3 ml/kg per h for 24 h or anuria for 12 h
***RIFLE*[13]**		
Risk	1.5-fold increase in serum creatinine level or >25% decrease in GFR	<0.5 ml/kg per h for 6 h
Injury	Twofold increase in serum creatinine level or >50% decrease in GFR	<0.5 ml/kg per h for 12 h
Failure	Threefold increase in serum creatinine level or 75% decrease in GFR or increase in serum creatinine level to ≥4 mg/dl (≥354 μmol/l) with an acute increase ≥0.5 mg/dl (≥44 μmol/l)	<0.3 mg/kg per h for 24 h or anuria for 12 h
Loss	Persistent AKI with complete loss of renal function (>4 weeks)	N/A
End-stage renal disease	ESRD (>3 months)	N/A

Figura 1.1: Tabela com os critérios de classificação para a AKI - Critérios KDIGO, AKIN e RIFLE (Leung et al., 2013).

No entanto, como referido acima, estes critérios de classificação da AKI baseiam-se nas alterações da creatinina sérica em relação ao valor da creatinina sérica basal. Este valor basal reflete a função renal antes de ser desenvolvida qualquer tipo de patologia. Normalmente, estes níveis basais não estão disponíveis ou não são conhecidos, uma vez que é preciso ter acesso ao historial clínico dos doentes. Assim, surge uma limitação na utilização destes critérios de classificação da AKI, que dificulta o seu diagnóstico (Siew et al., 2010).

Para ultrapassar este problema da indisponibilidade do valor da creatinina sérica basal, os clínicos utilizam outras opções de substituição deste valor. Por exemplo, o valor da SCr no ato da admissão hospitalar; o valor mínimo da SCr durante um determinado período de

internamento; valor estimado através de fórmulas de estimação, que determinam a GFR. No entanto, qualquer uma destas opções produz erros, levando a uma má classificação da AKI utilizando qualquer um dos critérios de classificação (Gaião and Cruz, 2010). A escolha da SCr basal tem um elevado efeito na prevalência da AKI, na gravidade da classificação da AKI e na mortalidade associada a esta patologia. Uma má classificação da AKI pode também levar a abordagens terapêuticas não adequadas (Siew et al., 2010).

A equação MDRD (*Modification of Diet in Renal Disease*) permite estimar a GFR, utilizando as variáveis creatinina sérica basal, sexo, idade e raça do indivíduo. Esta equação foi apresentada em 2002 por um grupo de investigadores na revista *American Journal of Kidney Diseases*, onde determinaram que esta equação é mais precisa para prever a GFR do que a *clearance* da creatinina urinária. A equação MDRD é dada pela seguinte expressão (Levey et al., 2015):

$$\begin{aligned} GFR =& 186 \times SCr(mg/dL)^{-1.154} \times idade^{-0.203}[\times 1.210 \text{ se raça negra}] \\ & [\times 0.742 \text{ se sexo feminino}]. \end{aligned} \tag{1.1}$$

Como é possível observar pela fórmula (1.1), quando se isola o valor da creatinina sérica obtém-se a partir da *back-calculation* o cálculo da sua estimação. Para o cálculo da creatinina sérica basal, fixa-se o valor da GFR em $75mL/min/1.73m^2$ (limite inferior da função renal considerada normal). O *Acute Dialysis Quality Initiative Working Group* (ADQI) recomenda que os clínicos utilizem a fórmula 1.2 no auxílio da classificação da AKI (Bagshaw et al., 2009).

$$SCrBasal = \frac{75}{186 \times idade^{-0.203}[\times 1.210 \text{ se raça negra}][\times 0.742 \text{ se sexo feminino}]}. \tag{1.2}$$

Embora este método de estimação seja muito utilizado em diversos estudos clínicos, ainda não foi validado (Bagshaw et al., 2009).

Závada et al. (2010) realizaram um estudo de aplicação da equação MDRD *versus* uma nova versão da MDRD criada pelo grupo e compararam os valores estimados com os valores reais conhecidos da SCr basal. O grupo concluiu que deve-se utilizar o valor da creatinina sérica basal conhecido, sempre que possível. No entanto, caso este valor não esteja disponível, o valor de creatinina sérica basal obtido pela MDRD pode ser utilizado como referência, embora contendo certas limitações (Závada et al., 2010).

Siew et al. (2010) realizaram um estudo de comparação dos erros introduzidos pelos diferentes métodos de estimação da SCr basal, utilizada para a classificação e prognóstico da lesão renal aguda. Foram comparados os valores da creatinina sérica no ato da admissão hospitalar, o menor valor da creatinina sérica num determinado período de internamento e o valor obtido através da equação MDRD. Os autores concluíram que a indisponibilidade do valor da creatinina sérica basal é um problema a ter em consideração na área da investigação da AKI. No entanto, a utilização de valores de substituição introduzem erros de medição, obtendo-se uma inadequada classificação e erros de prognóstico da AKI. Siew et al. (2010) recomendam novas investigações no desenvolvimento de novos métodos de estimação da SCr basal, para o

caso destes valores serem omissos ou não estejam disponíveis (Siew et al., 2010).

O diagnóstico da AKI mantém-se um problema devido à habitual inexistência do valor de SCr basal. Os clínicos continuam a usar a equação MDRD para extrapolar o seu valor, enquanto não houver outra ferramenta mais apropriada para a sua estimação ou enquanto não for utilizado outro biomarcador mais adequado (Pickering et al., 2009).

1.5 Objetivo

O objetivo deste estudo é o de encontrar um modelo com um bom poder preditivo que permita estimar o valor de creatinina sérica basal, através dos valores de um conjunto de variáveis obtidas a partir de uma amostra de doentes admitidos no Hospital Professor Doutor Fernando Fonseca (HFF) através do serviço de urgência.

O tipo de dados em análise, nomeadamente a distribuição contínua da variável resposta, justifica a utilização de várias abordagens de regressão. Assim sendo, o objetivo deste estudo passa pela aplicação de modelos lineares generalizados (MLG) e de modelos aditivos generalizados (MAG), bem como de modelos GAMLSS (*Generalized Additive Models for Location, Scale and Shape*).

Nesta análise, além da estimação do modelo são ainda verificadas as condições de aplicabilidade através da análise de resíduos.

Capítulo 2

Análise de Regressão - Conceitos Introdutórios

Neste capítulo será apresentado o conhecimento teórico dos modelos de regressão linear, modelos de regressão linear generalizados, os modelos de regressão aditivos generalizados e a transformação Box-Cox.

O principal objetivo da análise de regressão é sumariar a informação da amostra num modelo que analise a influência que uma ou mais variáveis independentes (ou covariáveis) têm sobre uma variável de interesse, a variável resposta (ou variável dependente). Outro objetivo da regressão é utilizar o modelo de regressão como modelo de predição, determinando a estimativa do valor expectável para a variável resposta através das suas covariáveis (Green and Silverman, 1994).

O modelo de regressão linear clássico foi introduzido por Legendre e Gauss no início do século XIX e dominou a estatística até ao século XX. No entanto, foram desenvolvidos outros novos modelos para ultrapassar as suas limitações. Em 1972, surgem os modelos lineares generalizados que, embora com uma estrutura linear e com a distribuição da variável resposta pertencente à família exponencial, se tornaram muito importantes na análise de dados (Turkman and Silva, 2000). No início da década de 90, Hastie e Tibshirani introduzem, pela primeira vez, o conceito de modelos aditivos generalizados, aplicando suavizadores aos modelos lineares generalizados (Stasinopoulos et al., 2015).

2.1 Modelo Linear

Considerando uma variável resposta Y e um conjunto de covariáveis, o modelo de regressão linear clássico é definido através da seguinte expressão (Silva, 2006):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

em que $\mathbf{X}$ representa a matriz das covariáveis com dimensão $n \times p$, sendo n o número total de indivíduos da amostra e p o número total das covariáveis. $\mathbf{Y}$ e $\boldsymbol{\epsilon}$ são vetores de dimensão $n \times 1$, sendo o último correspondente ao vetor dos erros aleatórios (Stasinopoulos and Rigby,

2014).

Dispondo de uma amostra de dimensão n, podemos escrever para o i-ésimo indivíduo o modelo linear com a seguinte expressão, $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij} + ... + \beta_p x_{ip} + \epsilon_i$, com $i = 1, 2, ..., n$ e $j = 1, 2, ..., p$, em que x_{ij} representa um elemento genérico da matriz de especificação $\mathbf{X}$. As variáveis aleatórias ϵ_i são independentes e identicamente distribuídas, com distribuição que se supõe $N(0, \sigma^2)$ (Stasinopoulos and Rigby, 2014). As condições de homocedasticidade e de erros não correlacionados são dadas por σ^2=var(ϵ_i), $\forall_i = 1, ..., n$, e cov(ϵ_i, ϵ_l) = 0, $i \neq l$, $l = 1, 2, ..., n$ (Silva, 2006). A distribuição da variável Y_i é normal com valor médio $\mu = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij} + ... + \beta_p x_{ip}$ e variância σ^2. Este modelo linear clássico apresenta uma relação linear entre o valor médio de Y e as covariáveis (Stasinopoulos et al., 2015), sendo o mais simples e o mais utilizado para análise de dados (Silva, 2006).

O vetor $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^\top$ da expressão 2.1, é composto pelos coeficientes de regressão desconhecidos do modelo de regressão linear. Estes coeficientes representam o peso que cada covariável $j = 1, ..., p$ dá à variável dependente do modelo (Hastie and Tibshirani, 1990). Os coeficientes β_j são estimados pelo método dos mínimos quadrados, minimizando a soma dos quadrados dos resíduos (Silva, 2006). A estimativa de $\boldsymbol{\beta}$ é dada pela seguinte equação, que corresponde ao estimador de máxima verosimilhança (Stasinopoulos et al., 2015):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \tag{2.2}$$

Desta forma os valores ajustados do modelo de regressão linear são obtidos pela expressão $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ (Stasinopoulos and Rigby, 2014).

2.2 Modelos Lineares Generalizados

Em 1972, Nelder e Weddernburn introduziram uma generalização ou extensão dos modelos lineares clássicos, apresentando os modelos lineares generalizados (MLG). Estes novos modelos ultrapassaram algumas limitações do modelo linear clássico. Nos MLG a distribuição normal da variável resposta Y é substituída pelas distribuições da família exponencial e a função de ligação $g(.)$ é utilizada para modelar a relação entre $\mu_i = E[Y_i|\mathbf{x}_i]$ e as variáveis independentes. A estimação do vetor β é feita de forma iterativa através do algoritmo de estimação de máxima verosimilhança (Stasinopoulos et al., 2015).

A variável Y tem obrigatoriamente de pertencer à família exponencial nos MLG. Alguns exemplos de distribuições da família exponencial são a distribuição normal, distribuição Poisson ou a distribuição Gamma. Diz-se que uma variável aleatória Y tem uma distribuição pertencente à família exponencial se a sua função densidade de probabilidade (f.d.p.) for definida pela seguinte forma (Silva, 2006):

$$f_Y(y|\theta; \phi) = \exp\left\{ \frac{y(\theta) - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \tag{2.3}$$

em que θ é a forma canónica do parâmetro de localização e ϕ é um parâmetro de dispersão que geralmente é considerado como conhecido. a(.), b(.) e c(.) são funções reais conhecidas

(Silva, 2006). O valor médio é dado por $E(Y) = \mu = b^{'}(\theta)$, enquanto a variância é dada por $var(Y) = a(\phi)b^{''}(\theta)$ (Stasinopoulos et al., 2015).

A estrutura dos modelos lineares generalizados são caracterizados por uma componente aleatória, uma componente sistemática e uma função de ligação a relacionar as duas componentes anteriores (Silva, 2006):

- Componente aleatória: dado o vetor de covariáveis $(x_{i1}, x_{i2}, ..., x_{ip})$, associado ao i-ésimo indivíduo, com $i = 1, ..., n$, as variáveis Y_i, além de terem uma distribuição pertencente à família exponencial, são (condicionalmente) independentes, $E[Y_i|\mathbf{x}_i] = \mu_i = b^{'}(\theta)$ (Silva, 2006).

- Componente sistemática (preditor linear): o valor esperado μ_i está relacionado com o preditor linear através da relação $\eta_i = g(\mu_i) = \beta_1 x_{i1} + ... + \beta_p x_{ip}$ ou por $\mu_i = h(\eta_i)$, em que $g(.) = h^{-1}(.)$ é uma função monótona e diferenciável, designada por função de ligação, que se supõe conhecida. A escolha desta função depende do tipo da distribuição da variável resposta. Quando o preditor linear coincide com o parâmetro canónico, $\eta_i = \theta_i$, a função de ligação correspondente chama-se de função de ligação canónica (Silva, 2006).

Caso a função de ligação escolhida seja a identidade e a distribuição da variável Y seja a normal, o modelo linear generalizado obtido é o corresponde ao modelo linear clássico (Wood, 2006).

2.3 Transformação Box-Cox

Quando para a variável dependente não se assume a distribuição normal, aplicam-se transformações a esta variável na tentativa de alcançar a normalidade procurada. Uma das transformações mais utilizadas é a transformação de Box-Cox, que aplica a seguinte função de transformação aos valores positivos de y (Florencio, 2010):

$$z = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & se \quad \lambda \neq 0, \\ log(y), & se \quad \lambda = 0, \end{cases} \tag{2.4}$$

em que o parâmetro λ, também denominado de parâmetro de transformação, é uma constante desconhecida que pode ser obtida por métodos computacionais. O programa R, através do *package* AID, possibilita a estimação do λ pela função **boxcoxnc**(). Esta função utiliza sete métodos diferentes de estimação, utilizando testes de normalidade, e um método artificial covariado, para a obtenção do parâmetro λ. Após a estimação de λ é aplicada à variável dependente a transformação z. Posteriormente é testada a normalidade dessa variável transformada, z, utilizando três testes de normalidade. Caso a normalidade não seja rejeitada, a transformação de Box-Cox foi de facto útil na solução do problema, obtendo-se a normalidade desejada para a variável dependente (Osborne, 2010).

2.4 Modelos Aditivos e Modelos Aditivos Generalizados

Os modelos aditivos generalizados são uma extensão dos modelos aditivos, introduzidos por Hastie e Tibshirani (1968 e 1987) e Stone (1986). Mais tarde, em 1990, os modelos aditivos generalizados foram descritos com mais detalhe por Hastie e Tibshirani (Schimek, 2000).

Os modelos de regressão linear clássicos ou os MLG são limitados devido à falta de flexibilidade da forma funcional que relaciona as variáveis independentes contínuas com a variável dependente (Silva, 2006). Os modelos aditivos generalizados (MAG) para ganharem maior flexibilidade, substituem a linearidade do modelo, mantendo, no entanto, a soma dos termos aditivos (Schimek, 2000). Uma vez que não é obrigatório impor uma relação linear para modelar a associação entre as covariáveis e a variável resposta, a limitação desta relação dos MLG é ultrapassada pelos MAG. Estes modelos recorrem a técnicas de suavização para encontrar a forma funcional que relaciona as variáveis independentes com a variável dependente (Silva, 2006).

Sempre que as funções suavizadoras seguirem certos requisitos de suavização, a sua estimação pode ser feita através de um *scatterplot smoother* de forma não paramétrica (Schimek, 2000). No entanto, a escolha destas suavizações é uma limitação dos MAG, uma vez que a escolha do grau de suavização pode ser diferente para cada uma das covariáveis (Silva, 2006).

O modelo aditivo é definido da seguinte forma (Hastie and Tibshirani, 1990):

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{2.5}$$

onde o erro ϵ é independente das variáveis X_j e tem $E(\epsilon) = 0$ e $Var(\epsilon) = \sigma^2$. As funções suavizadoras univariadas, f_j, $j = 1, ..., p$ são arbitrárias e suaves. Para evitar problemas de identificabilidade, a função $f_j(X_j)$ requere que $E[f_j(X_j)] = 0$, $\forall_j = 1, ..., p$ o que implica que $E(Y) = \alpha$, sendo α um parâmetro desconhecido (Hastie and Tibshirani, 1990).

O processo de estimação das funções univariadas f_j, e o mais utilizado para a estimação dos modelos aditivos, faz-se através do método de suavização iterativo, o algoritmo *backfitting*. A esperança condicional proporciona a motivação para este algoritmo, considerando que o modelo (2.5) seja o correto, temos para cada valor k, a relação $E[Y - \alpha - \sum_{j \neq k} f_j(X_j)|X_k] = f_k(X_k)$ (Hastie and Tibshirani, 1990). A ideia deste algoritmo, passa por ajustar o modelo, calcular os resíduos parciais resultantes e realizar novo ajuste até o ciclo convergir (Schimek, 2000).

Os MAG representam uma generalização dos MLG, uma vez que a extensão do modelo é obtida pela substituição do preditor linear βX pelo preditor aditivo $f(X)$. O método de estimação dos MAG também recorre ao algoritmo *backfitting* como nos modelos aditivos, no entanto, os MAG são mais flexíveis uma vez que têm uma função de ligação h conhecida (Silva, 2006).

Os MAG definem-se pela seguinte expressão:

$$\mu = h\left(\alpha + \sum_{j=1}^{p} f_j(X_j)\right) \tag{2.6}$$

onde $h(.)$ é uma função de ligação fixa e Y tem distribuição que também se assume pertencer à família exponencial, $f_j, j = 1, ..., p$ são funções suavizadoras designadas for funções parciais e α (Silva, 2006).

A estimação das funções f_j, $j = 1, ..., p$ nos MAG é feita através do algoritmo de *scores* local, semelhante ao algoritmo de *scores* de Fisher. O algoritmo de *scores* local é composto por dois ciclos, em que um deles corresponde ao algoritmo *backfitting* já referido anteriormente (Silva, 2006).

Para mais detalhes sobre modelos aditivos ou modelos aditivos generalizados consultar Hastie e Tibshirani, 1990.

Capítulo 3

Modelos GAMLSS - Generalized Additive Models for Location, Scale and Shape

As técnicas de regressão são muito utilizadas na análise de dados através do ajustamento de modelos que consideram uma variável resposta e uma ou mais covariáveis. No entanto, os métodos usuais têm algumas limitações, como por exemplo, a necessidade de assumir linearidade entre a variável dependente e as variáveis independentes, ou a distribuição da variável dependente ter de pertencer à família exponencial (Rigby and Stasinopoulos, 2009).

Os GAMLSS, introduzidos por Rigby e Stasinopoulos (2001, 2005), Stasinopoulos e Rigby (2007) e Akantziliotou et al. (2002), surgem na tentativa de aumentar a flexibilidade das técnicas de regressão, e de ultrapassar algumas limitações dos MLG, introduzidos por Nelder e Wedderburn em 1972, e dos MAG, introduzidos por Hastie e Tibshirani em 1990 (Rigby and Stasinopoulos, 2009).

Nos modelos GAMLSS todos os parâmetros da distribuição da variável resposta podem ser modelados em função das variáveis independentes, ou seja, não só é modelado o parâmetro de localização (ou posição - μ), mas também os parâmetros de escala (desvio padrão - σ) e os de forma (assimetria - ν e curtose - τ). Nos MLG e MAG, apenas é ajustado um modelo para o parâmetro μ, embora nas décadas de 1970-1980, já se tenha iniciado a modelação do parâmetro σ (Stasinopoulos et al., 2015). O ajustamento de modelos para cada parâmetro da distribuição da variável resposta, nos modelos GAMLSS, não tem, obrigatoriamente, que pertencer à família exponencial. Nestes modelos, a distribuição da variável dependente pode pertencer a um conjunto com cerca de 100 distribuições contínuas, discretas ou mistas. As distribuições pertencentes à família exponencial estão incluídas neste conjunto (Rigby and Stasinopoulos, 2009).

Diz-se que os modelos GAMLSS são semi-paramétricos porque implicam que a distribuição da variável dependente de quatro parâmetros, $D(\mu, \sigma, \nu, \tau)$, seja paramétrica, mas permite a utilização de funções suavizadoras não-paramétricas na modelação dos parâmetros da distribuição e nas funções das variáveis independentes (Rigby and Stasinopoulos, 2009)

(Spedicato et al., 2014).

Os modelos introduzidos neste capítulo permitem também ajustar modelos truncados, censurados ou modelos com misturas finitas de distribuições (Florencio, 2010). Também apresentam um bom desempenho em dados com elevada dispersão, com um elevado número de zeros para o caso de variáveis binárias, ou elevada assimetria ou curtose no caso de variáveis contínuas (Rigby and Stasinopoulos, 2009).

3.1 O Modelo

No modelo GAMLSS, as observações independentes da variável resposta Y, com $i = 1, 2, ..., n$, têm função densidade de probabilidade condicional, $f(y_i|\boldsymbol{\theta}^i)$ em que $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ representa o vetor de parâmetros da distribuição para cada Y_i. A notação destes modelos é dada por $Y_i|\boldsymbol{\theta}^i \sim D(\boldsymbol{\theta}^i)$, $Y_i|(\mu_i, \sigma_i, \nu_i, \tau_i) \sim D(\mu_i, \sigma_i, \nu_i, \tau_i)$, independentes para cada i, onde D representa a distribuição da variável resposta. Os dois primeiros parâmetros da distribuição, μ e σ, são os parâmetros de localização e escala, enquanto os restantes parâmetros, ν e τ, são considerados os parâmetros de forma (assimetria e curtose) (Rigby and Stasinopoulos, 2009).

Seja $\mathbf{y}^\top = (y_1, y_2, ..., y_n)$ o vetor dos valores observados da variável dependente e x_{kj}, com $k = 1, 2, 3, 4$ e $j = 1, 2, ..., J_k$, a variável independente inserida no modelo para o respetivo parâmetro da distribuição $\boldsymbol{\theta}_k$. Considerando $g_k(.)$ a função de ligação que relaciona cada k-ésimo parâmetro da distribuição com as variáveis independentes, é possível formular o seguinte modelo GAMLSS (Rigby et al., 2014):

$$\mathbf{Y} \sim D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}), \tag{3.1}$$

$$g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + s_{11}(x_{11}) + ... + s_{1J_1}(x_{1J_1}), \tag{3.2}$$

$$g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + s_{21}(x_{21}) + ... + s_{2J_2}(x_{2J_2}), \tag{3.3}$$

$$g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + s_{31}(x_{31}) + ... + s_{3J_3}(x_{3J_3}), \tag{3.4}$$

$$g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + s_{41}(x_{41}) + ... + s_{4J_4}(x_{4J_4}). \tag{3.5}$$

Este modelo é conhecido como modelo GAMLSS aditivo semi-paramétrico, onde $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$, e x_{kj} são vetores com dimensão n. $s_{kj}(.)$ representam as funções suavizadoras para as diferentes variáveis independentes, x_{kj} inseridas no modelo. $\mathbf{X}_k$ são matrizes fixas de ordem $n \times p_k$, que representam a parte linear do modelo, e os $\boldsymbol{\beta}_k$ representam os coeficientes de regressão linear (Stasinopoulos et al., 2015).

Os modelos GAMLSS que utilizam funções suavizadoras podem ser apresentados de outra forma. De facto, a matriz $s(x)$ pode ser substituída por $\mathbf{Z}\boldsymbol{\gamma}$, onde $\mathbf{Z}$ representa a matriz que depende apenas dos valores de x. $\boldsymbol{\gamma}$ representa um vetor de coeficientes a serem estimados, sujeito

a uma penalização quadrática na forma de $\lambda\boldsymbol{\gamma}^\top\mathbf{G}\boldsymbol{\gamma}$, onde $\mathbf{G} = \mathbf{D}^\top\mathbf{D}$ é a matriz de penalização e λ o valor que regula a quantidade de suavização, conhecido como parâmetro de suavização. Estas funções suavizadoras deverão ser referidas como funções suavizadoras penalizadas, que quando $\mathbf{Z}$ e $\mathbf{D}$ apresentam diferentes formulações, levam a diferentes tipos de funções suavizadoras com diferentes propriedades estatísticas. O modelo GAMLSS não-paramétrico pode, então, ser escrito da seguinte forma (Stasinopoulos et al., 2015):

$$\mathbf{Y} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}), \tag{3.6}$$

$$g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_{11}\boldsymbol{\gamma}_{11} + ... + \mathbf{Z}_{1k_1}\boldsymbol{\gamma}_{1J_1}, \tag{3.7}$$

$$g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_{21}\boldsymbol{\gamma}_{21} + ... + \mathbf{Z}_{2k_2}\boldsymbol{\gamma}_{2J_2}, \tag{3.8}$$

$$g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{Z}_{31}\boldsymbol{\gamma}_{31} + ... + \mathbf{Z}_{3k_3}\boldsymbol{\gamma}_{3J_3}, \tag{3.9}$$

$$g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{Z}_{41}\boldsymbol{\gamma}_{41} + ... + \mathbf{Z}_{4k_4}\boldsymbol{\gamma}_{4J_4}, \tag{3.10}$$

sujeito à penalização:

$$\sum_{k=1}^{4}\sum_{j=1}^{J_k}\lambda_{kj}\boldsymbol{\gamma}_{kj}^\top\mathbf{G}_{kj}\boldsymbol{\gamma}_{kj}. \tag{3.11}$$

3.2 Estimação do modelo

No caso de não existirem funções suavizadoras, o modelo reduz-se ao modelo GAMLSS paramétrico, $g_k(\boldsymbol{\theta}_k) = \mathbf{X}_k\boldsymbol{\beta}_k$, $k = 1, 2, 3, 4$ estimado pelo método de máxima verosimilhança, sendo a função de máxima verosimilhança dada por (Stasinopoulos et al., 2015):

$$\ell = \sum_{i=1}^{n} log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i), \tag{3.12}$$

em que f representa a f.d.p. da variável resposta. No entanto, para os modelos não-paramétricos é necessário recorrer ao método da máxima verosimilhança penalizada, através do logaritmo da função de verosimilhança dado por (Stasinopoulos et al., 2015):

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k}\lambda_{kj}\boldsymbol{\gamma}_{kj}^\top\mathbf{G}_{kj}\boldsymbol{\gamma}_{kj}. \tag{3.13}$$

O passo seguinte é estimar os vetores $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$, $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, ..., \gamma_{1J_1}, ..., \gamma_{41}, ..., \gamma_{4J_4})$ e $\boldsymbol{\lambda} = (\lambda_{11}..., \lambda_{1J_1}, ..., \lambda_{41}, ..., \lambda_{4J_4})$, através de algoritmos que permitem a maximização da função logaritmo da verosimilhança penalizada. Existem dois algoritmos básicos para atingir este objetivo que consideram valores fixos para o hiper-parâmetro λ, e que são os algoritmos RS e CG. Este último é uma generalização do algoritmo de Cole e Green (1992), que requer a primeira, a segunda e as derivadas cruzadas da função logaritmo de verosimilhança, em ordem aos quatro parâmetros da distribuição, $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$. O algoritmo RS é uma generalização do algoritmo de Rigby e Stasinopoulos (1996) e que, ao contrário do

CG, não necessita das derivadas cruzadas da função logaritmo da verosimilhança penalizada (Stasinopoulos et al., 2015).

3.2.1 Algoritmo RS

O algoritmo é descrito por dois ciclos, externo e interno, e pelo algoritmo *backfitting* modificado. Para atingir a convergência é necessário que todos eles sejam executados, ou seja, concluídos sem erros (Stasinopoulos et al., 2015).

Neste algoritmo apenas são necessários parâmetros de inicialização para $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4) = (\mu, \sigma, \nu, \tau)$. O algoritmo RS é bastante estável e consegue convergir de forma bastante rápida e eficiente, utilizando valores simples de inicialização como, por exemplo, valores constantes para os $\boldsymbol{\theta}, k = 1, 2, 3, 4$ (Stasinopoulos et al., 2015).

- Ciclo externo

Para o ciclo externo, ou iteração GAMLSS, primeiramente é necessário inicializar os parâmetros da distribuição com $\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \boldsymbol{\nu}_0$ e $\boldsymbol{\tau}_0$. O passo seguinte consta do ajustamento de um modelo para cada parâmetro da distribuição, dadas as últimas estimativas dos restantes parâmetros $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}})$. Segue-se então para o cálculo da *global deviance* ou desvio global $(\text{GD} = -2\hat{\ell})$ que, caso consiga convergir, o ciclo termina ou, caso contrário, é necessário a repetição do processo até ser obtida convergência (Stasinopoulos et al., 2015).

- Ciclo interno

O ciclo de iteração interno, ou iteração MLG, é um algoritmo de *scores* local muito semelhante ao utilizado no ajustamento dos MLG. A ideia deste algoritmo tem a ver com a repetição do ajustamento com a variável resposta modificada, utilizando uma ponderação, até ser atingida a convergência. Este ciclo é iniciado utilizando os resultados do ajustamento do ciclo anterior. Nos MLG este procedimento é conhecido como *Iterative Reweighted Least Squares* (Stasinopoulos et al., 2015).

A variável resposta modificada utilizada no ajuste dos modelos dos parâmetros θ_k e é dada por (Stasinopoulos et al., 2015):

$$\mathbf{z}_k = \boldsymbol{\eta}_k + \mathbf{w}_k^{-1} \bullet \mathbf{u}_k, \tag{3.14}$$

em que $\mathbf{z}_k, \boldsymbol{\eta}_k, \mathbf{w}_k$ e $\mathbf{u}_k$ são vetores de dimensão n e $\mathbf{w}_k^{-1} \bullet \mathbf{u}_k$ representa o produto de Hadamard de elemento por elemento dos vetores $(w_{k1}u_{k1}, w_{k2}u_{k2}, ..., w_{kn}u_{kn})^\top$. $\eta_k = g_k(\boldsymbol{\theta}_k)$ é o preditor dos k-ésimos parâmetros da distribuição. $\boldsymbol{u}_k$ é a função *score* correspondente obtida pela expressão (Stasinopoulos et al., 2015):

$$\mathbf{u}_k = \frac{\partial \ell}{\partial \boldsymbol{\eta}_k} = \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}_k}\right) \bullet \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right), \tag{3.15}$$

em que $\mathbf{w}_k$ é o vetor dos pesos dos parâmetros da distribuição utilizado nas iterações e definido por (Stasinopoulos et al., 2015):

$$\mathbf{w}_k = -\mathbf{f}_k \bullet \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right) \bullet \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right). \tag{3.16}$$

A definição de $\mathbf{f}_k$ pode ser feita de três formas diferentes, dependendo da informação disponível acerca da distribuição. Assim $\mathbf{f}_k$ pode ser definido por: $E\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_k^2}\right]$, levando à utilização do algoritmo do *score* de Fisher; $\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_k^2}$, conduzindo à utilização do algoritmo de Newton-Raphson; $\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}_k}\right) \bullet \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}_k}\right)$, para a utilização do algoritmo *quasi* Newton-Raphson (Stasinopoulos et al., 2015).

Através dos parâmetros estimados pelo ciclo externo, $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$, são calculados os pesos $\mathbf{w}_k$ e por sua vez as variáveis $\mathbf{z}_k$ que serão utilizados na iteração seguinte. O ciclo continua até o valor da GD não se alterar no fim de cada ciclo (Stasinopoulos et al., 2015).

- Algoritmo de *backfitting* modificado

As estimações de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são determinadas através deste algoritmo. O algoritmo de *backfitting* modificado ajusta variáveis independentes e suavizadores a $\mathbf{z}_k$, utilizando $\mathbf{w}_k$ através do *backfitting*, para atualizar as estimativas do parâmetro θ_k (Stasinopoulos et al., 2015).

São necessários (dentro do ciclo interno) os valores de $\mathbf{w}_k$ e $\mathbf{z}_k$ e a imputação do(s) valor(es) do(s) coeficiente(s) do(s) suavizador(es) $\boldsymbol{\gamma}$, para calcular os resíduos parciais ($\boldsymbol{\epsilon}_k = \mathbf{z}_k - \boldsymbol{\gamma}_k$) dos $\boldsymbol{\beta}$. Posteriormente, ajusta-se um modelo *Weighted Least Squares* (WLS) aos resíduos para obter novas estimativas de $\hat{\beta}_k$. O passo seguinte é obter os resíduos parciais de $\boldsymbol{\gamma}$, ($\boldsymbol{\epsilon}_k = \mathbf{z}_k - \mathbf{X}\boldsymbol{\beta}_k$) em ordem a cada suavizador e ajustar um modelo *Penalised Weighted Least Squares* (PWLS) para obter novos valores de $\hat{\boldsymbol{\gamma}}_k$. Este processo é repetido até os valores de $\hat{\boldsymbol{\beta}}_k$ e $\hat{\boldsymbol{\gamma}}_k$ não se alterarem mais (Stasinopoulos et al., 2015) e (Rigby et al., 2014).

Após estes processos, são atualizados os valores de $\boldsymbol{\theta}_k$ e de $\boldsymbol{\eta}_k$ para serem utilizados no ciclo interno e posteriormente para calcular a *global deviance* do ciclo externo, até finalmente o modelo RG convergir (Rigby et al., 2014).

Para mais detalhes sobre o algoritmo RS consultar Stasinopoulos et al. (2015).

Ao contrário dos modelos **gam**() disponíveis no R, os modelos **gamlss**() não ajustam a componente linear e suavizadora simultaneamente. Embora o algoritmo do **gam**() seja mais rápido e consiga ajustar suavizadores penalizados, o algoritmo do **gamlss**() permite o ajustamento de outros suavizadores não contemplados pelos **gam**() como, por exemplo, o *loess*, os *splines* cúbicos ou as redes neuronais (Stasinopoulos et al., 2015).

3.2.2 Algoritmo CG

Tanto o algoritmo CG como o RS, conseguem estimar os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, dado um vetor de hiper-parâmetros $\boldsymbol{\lambda}$, de maximização do logaritmo da função verosimilhança penalizada. Uma vez que este algoritmo necessita da derivada cruzada da função de verosimilhança, este

consegue atualizar os valores estimados dos parâmetros da distribuição em simultâneo. No entanto, embora pareça o algoritmo mais adequado, é instável e tem dificuldades em convergir. O algoritmo RS é mais estável e, na maioria dos casos, consegue atingir a convergência mais rapidamente do que o CG. Assim, o algoritmo utilizado, por defeito, na função **gamlss**() do *package* GAMLSS do programa R é o algoritmo RS (Stasinopoulos et al., 2015).

Neste estudo não será apresentado o algoritmo CG. Para mais detalhes consultar Stasinopoulos et al. (2015).

3.2.3 Estimação de λ

Para ambos os algoritmos RS e CG, a estimação dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, através da função de máxima verosimilhança penalizada era feita através da fixação do hiper-parâmetro $\boldsymbol{\lambda}$. No entanto, a estimação dos hiper-parâmetros $\boldsymbol{\lambda}$ pode ser feita de forma local (dentro dos algoritmos RS ou CG, ou seja, no algoritmo *backfitting*) ou de forma global (fora dos algoritmos RS ou CG), através de três diferentes metodologias: GCV (*Generalised cross validation* - validação cruzada generalizada), GAIC (*Generalized Akaike Information Criterion*) ou por métodos baseados na máxima verosimilhança (Stasinopoulos et al., 2015).

Para mais informações sobre estas metodologias locais consultar Stasinopoulos et al. (2015).

Os autores Stasinopoulos et al. (2015) recomendam a utilização dos métodos locais uma vez que são mais rápidos e normalmente produzem resultados semelhantes aos métodos globais. O *package* GAMLSS do programa R possibilita a estimação de suavizadores através do método GAIC, de forma global, pela utilização da função **find.hyper**(). A função **find.hyper**() possibilita a estimação do hiper-parâmetro de suavização, obtendo bons resultados na procura de ótimos graus de liberdade para o suavizador. Esta função utiliza a função **optim**() do programa R para minimizar o critério GAIC, para escolher o hiper-parâmetro $\boldsymbol{\lambda}$. No entanto, para grandes bases de dados esta função pode tornar-se demorada em comparação com a utilização de métodos locais. Desta forma é aconselhada a utilização dos métodos locais em substituição da utilização da função **find.hyper**() (Stasinopoulos et al., 2015).

3.2.4 Considerações gerais dos algoritmos

O algoritmo RS é utilizado para estimar $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, recorrendo à verosimilhança penalizada, para valores de $\boldsymbol{\lambda}$ fixos. No entanto, é possível estimar o hiper-parâmetro $\boldsymbol{\lambda}$, por métodos locais ou globais. O método local estima cada λ_{kj} por cada iteração do algoritmo *backfitting*, enquanto o método global é aplicado fora do algoritmo RS. Por vezes, os métodos de estimação locais para o parâmetro de suavização, podem tornar os algoritmos RS e CG mais instáveis provocando um aumento do desvio global (Stasinopoulos et al., 2015).

3.3 Termos aditivos do modelo

Os modelos GAMLSS permitem modelar todos os parâmetros da distribuição pelas covariáveis através de relações na forma linear e/ou não-linear e/ou através de funções suavizadoras

não-paramétricas. Uma relação não-linear pode ser 'paramétrica não-linear' ou um suavizador. Estas relações afetam cada preditor de cada parâmetro da distribuição, resultando na alteração da forma da distribuição da variável dependente (Stasinopoulos et al., 2015).

3.3.1 Relação linear paramétrica

A relação linear considerada nos modelos GAMLSS é semelhante à dos modelos lineares e dos modelos lineares generalizados. Quando não se verifica a linearidade da relação entre a variável resposta e as covariáveis é comum transformar estas últimas utilizando polinómios ou ainda através de técnicas de suavização como, por exemplo, os *splines* (Silva, 2012).

3.3.2 Relação não-linear paramétrica

Um dos exemplos mais simples desta relação são os polinómios aplicados às variáveis independentes que conferem uma certa flexibilidade à curva de regressão através da potência definida do polinómio. Existem diferentes tipos de polinómios como os ortogonais, *fractional*, *piecewise* e *B-splines* (Stasinopoulos et al., 2015).

Os polinómios *piecewise* são uma importante ferramenta para a modelação na forma penalizada, uma vez que estes polinómios são muito conhecidos como técnicas de suavização não-paramétricas. Estes polinómios são utilizados quando existe uma alteração da relação entre a variável dependente e a variável independente. O nome '*splines*' é aplicado aos polinómios *piecewise* (Stasinopoulos et al., 2015).

3.3.3 Suavizadores

Uma importante propriedade dos suavizadores é não assumirem a forma paramétrica que relaciona a variável resposta com as covariáveis. Os suavizadores permitem que o dados determinem qual essa relação funcional. No entanto, o conceito 'não-paramétrico' pode ser mal interpretado, uma vez que os suavizadores também permitem estimar parâmetros, como o de suavização $\boldsymbol{\lambda}$, que determina a quantidade de suavização necessária aos dados. A escolha do parâmetro de suavização é um fator importante em qualquer técnica de suavização (Stasinopoulos et al., 2015).

Existem diversas funções suavizadoras disponíveis no *package* GAMLSS que são divididas em duas categorias, suavizadores penalizados e os restantes (Figura 3.1). Os suavizadores penalizados utilizam a penalização quadrática para controlar a quantidade de suavização, e os restantes suavizadores utilizam penalizações não quadráticas, para obter a função suavizadora. Alguns exemplos destas funções são *P-splines e cubic splines, loess curve fitting, neural networks*, entre outros (Stasinopoulos et al., 2015).

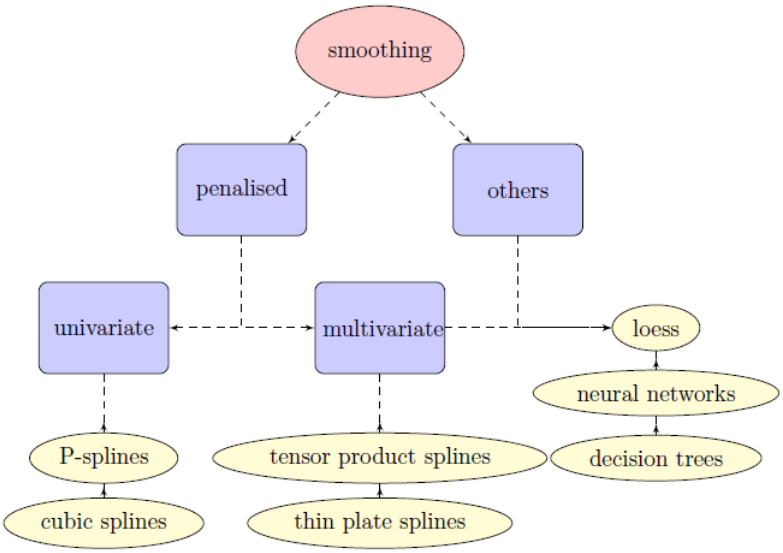

Figura 3.1: Funções suavizadoras disponíveis no *package* GAMLSS (Stasinopoulos et al., 2015).

Neste capítulo serão só referidos os suavizadores penalizados univariados: *P-splines* e os suavizadores cúbicos (*cubic splines*).

Um suavizador simples univariado é uma generalização da regressão linear simples e pode ser representado da seguinte forma (Stasinopoulos et al., 2015):

$$E(Y_i) = s(x_i), \tag{3.17}$$

para $i = 1, ..., n$, onde Y é a variável resposta contínua, com distribuição normal (μ_i, σ^2). $s(.)$ é uma função suavizadora definida pela relação $\mu_i = s(x_i)$, uma vez que se quer estimar $s(x_i)$, sendo x_i a covariável. A função $s()$ tem de assumir algumas prioridades que irão definir o tipo de suavizador a ser aplicado, como por exemplo para os *cubic splines* a primeira e segunda derivada da função têm de ser contínua (Stasinopoulos et al., 2015).

Uma vantagem da utilização de funções suavizadoras, em vez da utilização de um ajuste paramétrico é o facto de os suavizadores terem um comportamento local. Os suavizadores são influenciados por observações locais, em vez de observações mais distantes. Este comportamento pode ser representado pela introdução dos modelos locais de regressão, os suavizadores de regressão local.

• Suavizadores de regressão local

Um suavizador local utiliza apenas um subgrupo da amostra, em vez de serem usados a totalidade dos dados, para o cálculo do valor suavizado. O subgrupo é determinado pela janela que corresponde a um intervalo de valores de uma das covariáveis, permitindo que apenas os que estão dentro da janela sejam utilizados para o cálculo do valor suavizado. Os suavizadores de regressão local incluem três passos importantes: a escolha da janela de suavização, a escolha do grau dos polinómios a aplicar e a forma como se realizará a estimação dos valores. Este último passo é definido pelo tipo de polinómios a utilizar, *weighted* ou *unweighted*. A regressão

local polinomial *weighted* requer a definição de uma função Kernel e do seu hiper-parâmetro. Para a regressão local polinomial *unweighted* a janela é determinada pelo *span*, $(2k+1)/n$, onde k é o número de observações à direita/esquerda do ponto central do subgrupo e n o número total de observações. O *span* pode variar entre 0 e 2. Para o caso de ser muito próximo de 0, então a janela irá conter apenas uma observação, se próximo de 2 então irá conter todas as observações. Para a regressão local *unweighted*, o *span* é considerado o parâmetro de suavização λ. A escolha do parâmetro de suavização, λ é um dos tópicos mais importantes das técnicas de suavização (Stasinopoulos et al., 2015).

• Suavizadores penalizados univariados

Os suavizadores penalizados univariados são os mais importantes do *package* GAMLSS devido à sua flexibilidade e à possibilidade de serem aplicados em diversas situações. A solução dos suavizadores penalizados é a solução da minimização da quantidade Q (3.18), em ordem a $\boldsymbol{\gamma}$ (Stasinopoulos et al., 2015):

$$Q = (y - \mathbf{Z}\boldsymbol{\gamma})^\top \mathbf{W}(y - \mathbf{Z}) + (\lambda \boldsymbol{\gamma}^\top \mathbf{G}\boldsymbol{\gamma}), \tag{3.18}$$

sendo $\mathbf{Z}$ a matriz de dimensão $n \times p$, já definida nos modelos GAMLSS (3.7, 3.8, 3.9 e 3.10), γ o vetor de parâmetros de dimensão p a serem estimados, $\mathbf{W}$ a diagonal da matriz dos pesos de dimensão $n \times n$, $\mathbf{G}$ a matriz de penalização $(p \times p)$, λ o parâmetro suavizador e y a variável resposta.

A solução do problema de minimização de 3.18 é dada pela seguinte expressão:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{G})^{-1} \mathbf{Z}^\top \mathbf{W} y. \tag{3.19}$$

Diferentes $\mathbf{Z}$ e $\mathbf{G}$ produzem diferentes suavizadores e $\mathbf{W}$ é utilizado no algoritmo *backfitting* do modelo GAMLSS. Os valores ajustados obtidos são dados por (Stasinopoulos et al., 2015):

$$\hat{y} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{G})^{-1} \mathbf{Z}^\top \mathbf{W} y = \mathbf{S} y, \tag{3.20}$$

onde $\mathbf{S}$ representa a matriz de suavização. Outra quantidade de interesse dos GAMLSS é o traço da matriz $\mathbf{S}$, utilizada para determinar os graus de liberdade do suavizador, $tr(\mathbf{S}) = tr[\mathbf{Z}(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{G})^{-1} \mathbf{Z}^\top \mathbf{W}]$ (Stasinopoulos et al., 2015).

Existem diferentes métodos de estimação dos suavizadores de forma automática: através da GCV, do critério GAIC ou do método de máxima verosimilhança (Stasinopoulos et al., 2015). Para mais informações relativas a estes métodos consultar Stasinopoulos et al. (2015).

• *P-splines*

A função mais comum dos *P-splines* é a **pb**(), *Penalised B-splines*, que utiliza a quantidade Q (3.18), substituindo $\mathbf{Z}$ por $\mathbf{B}$ correspondente a um B-*spline* de base de um polinómio *piecewise* de grau d, na equação 3.19. O coeficiente γ é penalizado consoante a matriz de penalização $\mathbf{G} = \mathbf{D}_k^\top \mathbf{D}_k$ adequada ao parâmetro θ_k da distribuição (Stasinopoulos et al., 2015).

Alguns argumentos da função suavizadora **pb**() disponível no *package* GAMLSS são: x - variável independente; *df* - número de graus de liberdade; *lambda* - parâmetro suavizador. Se o número de graus de liberdade e o lambda não forem especificados, a função estima valores pelo método local de máxima verosimilhança (Stasinopoulos et al., 2015). Existem algumas variações dos P-*splines* que podem ser consultados com mais detalhe em Stasinopoulos et al. (2015).

- *Cubic splines*

Para os suavizadores *cubic splines* as funções utilizadas são **cs**() e **scs**() disponíveis no *package* GAMLSS, as quais são uma variação da função **smooth.spline**(). Existem diferenças entre os suavizadores *P-splines* e os *cubic splines*. As funções ajustadas para ambos os suavizadores são habitualmente muito semelhantes, no entanto os *P-splines* ajustam a função suavizadora pela penalização do parâmetro γ e os *cubic splines* pela penalização da segunda derivada da função de verosimilhança. **cs**() e **scs**() têm diferentes valores por defeito quando *df* e *lambda* não são especificados. **cs**() utiliza 3 graus de liberdade extra para o suavizador, enquanto que **scs**() estima λ por defeito e, por sua vez, o número de graus de liberdade automaticamente pelo método GCV. No entanto, um suavizador que utiliza o método GCV, pode ser mais instável do que aquele com valores fixos para os graus de liberdade, especialmente para amostras de dimensões pequenas (Stasinopoulos et al., 2015).

No entanto, os graus de liberdade para a função **scs**() podem ser estimados, manualmente pela função **find.hyper**() (Stasinopoulos et al., 2015).

Quando são utilizados suavizadores nos modelos GAMLSS, é preciso ter em atenção à análise do *output* obtido pelo programa R. O *output* do modelo GAMLSS decompõe o suavizador na sua parte 'linear' e parte 'não-linear', apresentando apenas o coeficiente e erro padrão da parte 'não-linear' (Stasinopoulos et al., 2015).

Para mais informações sobre suavizadores consultar Stasinopoulos et al. (2015).

3.4 Seleção do Modelo

O *package* GAMLSS tem várias técnicas e funções que possibilitam a seleção do modelo estatístico mais adequado aos dados em estudo. Aquele que melhor se ajustar aos dados deverá ter em conta o objetivo do estudo, e ainda, proporcionar um balanço entre a sobrestimação (*overfitted*) e a subestimação (*underfitted*) que, em termos de inferência estatística, significa um equilíbrio entre a variância e o viés dos estimadores. Um modelo sobrestimado não é um modelo aconselhado para a predição, devido à sua elevada variância. Os modelos subestimados são muito enviesados, mas com variância pequena que, por vezes, podem apresentar uma melhor predição dos valores estimados (Stasinopoulos et al., 2015).

Seja $\mathcal{M}$ o modelo estatístico ajustado aos dados e o respetivo desvio global, (GD= $-2\ell(\hat{\theta})$). Podem-se distinguir dois modelos, no que diz respeito ao seu desempenho, $\mathcal{M}_0$ e $\mathcal{M}_1$, através de GD_0 e GD_1 e do número de graus de liberdade df_0 e df_1, respetivamente. O $\mathcal{M}_0$ representa o

modelo mais simples e é considerado uma subclasse do $\mathcal{M}_1$, uma vez que este é o mais completo. O modelo $\mathcal{M}_0$ pode ser referido como o modelo aninhado no modelo $\mathcal{M}_1$. Ambos os modelos podem ser comparados utilizando a estatística de teste da razão de verosimilhanças generalizadas (Generalized likelihood ratio test statistic), Λ (Stasinopoulos et al., 2015).

$$\Lambda = \mathrm{GD}_0 - \mathrm{GD}_1. \tag{3.21}$$

Esta estatística tem uma distribuição assintótica χ^2, sob a hipótese nula de que o modelo correto é o $\mathcal{M}_0$, com graus de liberdade $d = df_0 - df_1$. Para modelos com termos aditivos não-paramétricos é possível utilizar o mesmo teste, mas é necessário ter em conta que o número de graus de liberdade é dado pelo traço da matriz suavizadora resultante do ajustamento (Stasinopoulos et al., 2015).

No caso de modelos GAMLSS não aninhados, é possível compará-los pelo critério GAIC, penalizando a sobrestimação, adicionando ao desvio global do modelo, uma $k = 1, 2, 3, 4$ penalização por cada grau de liberdade do modelo. $\mathrm{GAIC}(k) = \mathrm{GD} + (k \times df)$, df representa o número de graus de liberdade utilizado no modelo. O modelo com menor valor de GAIC é o selecionado como melhor modelo ajustado aos dados. O critério AIC é um caso especial do GAIC, onde $k = 2$. Este critério, em casos práticos, leva à seleção de modelos sobreajustados. (Stasinopoulos et al., 2015).

Abaixo encontram-se os quatro passos para a seleção das quatro componentes de um modelo GAMLSS, $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$:

1. $\mathcal{D}$ representa a distribuição da variável resposta,
2. $\mathcal{G}$ representa o conjunto das funções de ligação,
3. $\mathcal{T}$ especifica as variáveis independentes incluídas nos modelos dos parâmetros da distribuição μ, σ, ν e τ,
4. Λ especifica os hiper-parâmetros suavizadores que determinam a quantidade de suavização $\mathrm{s}_{kj}()$.

Na construção de um modelo GAMLSS, as componentes acima devem ser especificadas da forma mais objetiva possível (Stasinopoulos et al., 2015).

3.4.1 Componente $\mathcal{D}$ - Seleção da distribuição

A seleção da distribuição da variável dependente é feita através do critério GAIC(k). São ajustados e comparados diferentes modelos GAMLSS com diferentes distribuições e é selecionado aquele com menor valor de GAIC. (Stasinopoulos et al., 2015).

A função utilizada para escolher a distribuição da variável resposta, disponível no *package* GAMLSS, é a **fitDist**(). Esta função utiliza a função **gamlss**() para ajustar diferentes distribuições à variável de interesse. Os argumentos da função **fitDist**() são o vetor dos valores da variável resposta, o valor da penalização, k, do critério GAIC, por defeito $k = 2$, tipo de distribuições a ajustar, *type*. Este último argumento é definido por '*realline*', '*realplus*' ou

'*realAll*', os quais definem grupos de distribuições a ajustar pela função **fitDist**(). Para o caso de uma determinada função não se encontrar em nenhum dos grupos descritos acima, podem ser adicionadas pelo argumento *extra*. A seleção da distribuição é feita pelo critério GAIC(k) e estes valores podem ser visualizados pelos *fits* e *failed* obtidos pela função **fitDist**() (Stasinopoulos et al., 2015).

Uma vez escolhida a distribuição é possível visualizar, através do programa R, a distribuição ajustada à variável dependente, pela função **histDist**(). Esta obtém valores constantes para os parâmetros da distribuição. A função necessita apenas como argumentos a variável Y e a especificação da distribuição escolhida pela função **fitDist**() (Stasinopoulos et al., 2015).

3.4.2 Componente $\mathcal{G}$ - Seleção das funções de ligação

A seleção das funções de ligação para cada parâmetro da distribuição é usualmente determinada pela distribuição escolhida para a variável resposta. Cada distribuição já tem as respetivas funções de ligação selecionadas para cada parâmetro da distribuição (Stasinopoulos et al., 2015).

3.4.3 Componente $\mathcal{T}$ - Seleção dos termos aditivos

Os termos aditivos a serem inseridos no modelo para cada parâmetro da distribuição θ_k, $k = 1, 2, 3, 4$, podem ser lineares como suavizadores. Para a respetiva distribuição da variável resposta, a seleção dos termos aditivos tem de ser feita para todos os parâmetros da distribuição. Os termos adicionados influenciam os parâmetros da distribuição de forma diferente. O *package* GAMLSS disponibiliza três diferentes métodos de seleção de termos aditivos **stepGAIC**(), **stepGAICAll.A**() e **stepGAICAll.B**() (Stasinopoulos et al., 2015).

Existem diferentes funções que o *package* GAMLSS disponibiliza para selecionar as covariáveis. As funções básicas são **addterm**() e **dropterm**() que permitem a adição e a remoção de covariáveis na estimação de determinado parâmetro de distribuição. Estas funções são utilizadas na função **stepGAIC**(), que seleciona os termos aditivos para cada parâmetro μ, σ, ν e τ, individualmente. Para além da função anterior, existem as funções **stepGAICAll.A**() e **stepGAICAll.B**(), que selecionam os termos aditivos para todos os parâmetros da distribuição em simultâneo (Stasinopoulos et al., 2015).

A função **stepGAIC**() faz uma seleção *stepwise* do modelo recorrendo ao critério GAIC. Esta função é baseada na função **stepAIC**() dada pelo *package* MASS. A seleção das covariáveis é feita para qualquer um dos parâmetros da distribuição de interesse, uma vez que existe um argumento *what* que pode conter um dos seguintes valores "*mu*", "*sigma*", "*nu*"ou "*tau*", correspondente a μ, σ, ν e τ. Para iniciar a função é necessário indicar o modelo inicial e o modelo mais completo. A seleção das covariáveis é feita pelo método *stepwise* (*both, backward* ou *forward*) (Stasinopoulos et al., 2015).

A função **stepGAICAll.A**() também seleciona covariáveis utilizando o critério GAIC, semelhante à função anterior, mas com outra estratégia de seleção. Primeiro faz a seleção *forward* GAIC das covariáveis para o parâmetro μ, considerando constantes σ, ν e τ. No passo seguinte, realiza o mesmo procedimento para o parâmetro seguinte, σ, considerando ν e τ

constantes, mas com μ já ajustado com as covariáveis selecionadas no passo anterior. Após fazer o mesmo para ν e τ, a função aplica a seleção *backward* GAIC ao parâmetro ν, mantendo as covariáveis selecionadas nos passos anteriores. Realiza-se o mesmo procedimento para σ e μ, sucessivamente, sempre mantendo as covariáveis selecionadas para os parâmetros dos passos anteriores. O modelo final irá conter uma sub-seleção das covariáveis, não necessariamente igual, para cada parâmetro da distribuição (Stasinopoulos et al., 2015).

A segunda estratégia de seleção, **stepGAICAll.B**(), obriga a que todos os parâmetros da distribuição tenham as mesmas covariáveis, ou seja, quando uma covariável é inserida num modelo é inserida para todos os restantes parâmetros do modelo. A inclusão das covariáveis no modelo pode ser feita na direção *forward*, *both* ou *backward*, utilizando sempre o critério GAIC. Esta função tem os mesmos argumentos que a função **stepGAICAll.A**(), exceto o argumento *what* que, neste caso, não é necessário, uma vez que se incluem sempre as mesmas covariáveis para todos os parâmetros. Para esta função também é necessário especificar o modelo mais simples e o modelo mais complexo (Stasinopoulos et al., 2015).

3.4.4 Componente Λ - Seleção dos parâmetros de suavização

Cada suavizador incluído no modelo, para qualquer parâmetro da distribuição, tem pelo menos um hiper-parâmetro de suavização, λ, associado. O valor de λ pode ser estimado ou fixo. A forma tradicional de fixação do hiper-parâmetro é feita fixando do número de graus de liberdade como sugerido por Hastie e Tibshirani (1990). No entanto, é desejável estimá-lo. O *package* GAMLSS consegue fazer a estimação de λ automaticamente através dos três métodos de estimação já referidos: GCV, GAIC e método de máxima verosimilhança. Os autores aconselham o método local devido à sua rapidez e também porque consegue obter resultados semelhantes ao método global (Stasinopoulos et al., 2015).

3.5 Diagnóstico do modelo

Uma análise importante utilizada para determinar se o modelo GAMLSS escolhido tem um bom ajustamento aos dados é a análise de resíduos. Os resíduos utilizados nos modelos de regressão linear simples, $y_i = \beta_0 + \beta_1 x_i + e_i$, calculam-se pela diferença entre o valor observado e o valor ajustado, $\hat{\varepsilon}_i = y_i - \hat{y}_i$, enquanto que os padronizados são definidos por $(y_i - \hat{y}_i)/\hat{\sigma}\sqrt{(1-h_{ii})}$, onde h_{ii} representa o elemento ii da diagonal da matriz *hat* dos resíduos obtidos pelos MLG. Ainda é possível calcular os desvios residuais e os de *Pearson*. Os primeiros não são bem definidos para distribuições da variável resposta Y com múltiplos parâmetros, e os de *Pearson* não são normalizados e não são apropriados para dados com elevada assimetria e curtose (Stasinopoulos et al., 2015).

Os resíduos recomendados para os modelos GAMLSS são os *normalised randomised quantile residuals*. O cálculo destes resíduos para uma variável de interesse contínua é diferente do cálculo para uma variável discreta. Uma vez que neste estudo a variável de interesse é contínua, apenas irá ser abordado o cálculo dos *normalised randomised quantile residuals* para este tipo

de variável (Stasinopoulos et al., 2015).

Para mais detalhes em relação ao cálculo dos resíduos *normalised randomised quantile residuals* para variáveis contínuas consultar Stasinopoulos et al. (2015).

Na próxima secção são introduzidos os *normalised randomised quantile residuals* e as respetivas funções que possibilitam a construção dos gráficos utilizados para os analisar.

3.5.1 Normalised Randomise Quantile Residuals

A vantagem destes resíduos tem a ver com o facto de que, independentemente da distribuição selecionada para a variável resposta, os *normalised randomised quantile residuals* terão de ter uma distribuição normal, quando o modelo GAMLSS ajustado for adequado aos dados. Os resíduos dos GAMLSS são calculados pela seguinte fórmula: $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, onde Φ^{-1} é a inversa da função de distribuição cumulativa da N(0,1) e $\hat{u}_i$ representa os quantis dos resíduos. Sendo y uma observação de uma variável aleatória contínua, os $\hat{u} = F(y|\hat{\theta})$. Se o modelo GAMLSS for o correto, $\hat{u}$ terá uma distribuição uniforme entre 0 e 1. Para concluir, $\hat{r} = \Phi^{-1}(\hat{u}) = \Phi^{-1}[F(y|\hat{\theta})]$ e $\hat{r}$ tem, aproximadamente, uma distribuição normal padronizada (Stasinopoulos et al., 2015).

Através da função **residuals**() do programa R é possível obter o vetor dos resíduos do modelo GAMLSS. Para analisar a normalidade utilizam-se métodos gráficos, como o *QQ-plot*, resíduos *versus* o índice da observação, resíduos *versus* valores ajustados $\hat{y}$, gráfico densidade de Kernel ou o *Worm plot* (Stasinopoulos et al., 2015).

3.5.2 *Worm plot*

Os *Worm plots* dos resíduos foram introduzidos por Buuren and Fredriks em 2001, para identificar regiões em que o modelo não é bem ajustado aos dados. Através do *Worm plot* é possível observar pontos (a amarelo), que indicam o quão afastados estão os resíduos do seu valor expectável, representado pela linha a vermelho a tracejado. Se o modelo ajustado for correto, será de esperar que 95% dos pontos devem estar entre os dois semi-círculos desenhados a tracejado. Uma elevada percentagem de pontos dentro dos semi-círculos indicam que a distribuição ajustada ou as variáveis selecionadas para o modelo podem não ser as mais corretas. A curva tracejada a vermelho é uma curva de ajuste cúbica aos pontos, e a forma obtida reflete as diferentes inadequabilidades do modelo (Figura 3.2) (Stasinopoulos et al., 2015).

Na tabela 3.1 encontram-se vários tipos de *Worm plot* de modelos GAMLSS mal ajustados. Na figura 3.2 estão representadas as formas gráficas desses modelos.

Tabela 3.1: Forma dos *Worm plot* de modelos GAMLSS mal ajustados (Stasinopoulos et al., 2015)

Forma do *Worm plot*	**Resíduos**	**Variável dependente**
Acima da curva de ajustamento	Valor médio elevada	parâmetro de localização muito baixo
Abaixo da curva de ajustamento	Valor médio baixa	parâmetro de localização muito elevado
Declive positivo (+)	Variância elevada	parâmetro de escala muito baixo
Declive negativo (-)	Variância baixa	parâmetro de escala muito elevado
Forma 'U'	Assimetria (+)	assimetria muito baixa
Forma 'U' invertida	Assimetria (-)	assimetria muito elevada
Forma 'S' com *left bent down*	*lepto-kurtosis*	curtose baixa
Forma 'S' com *left bent up*	*platy-kurtosis*	curtose elevada

Os gráficos (a) e (b) (Figura 3.2) mostram *Worm plots* quando o modelo do parâmetro de localização tem um ajustamento inadequado aos dados, uma vez que os resíduos se encontram todos acima ou abaixo da reta $y = 0$. Os *Worm plots* (c) e (d) representam um mau ajustamento do modelo para o parâmetro de escala. Já os gráficos (e) e (f) indicam falha do ajustamento do parâmetro de assimetria e (g) e (h) falha do ajustamento do parâmetro de curtose.

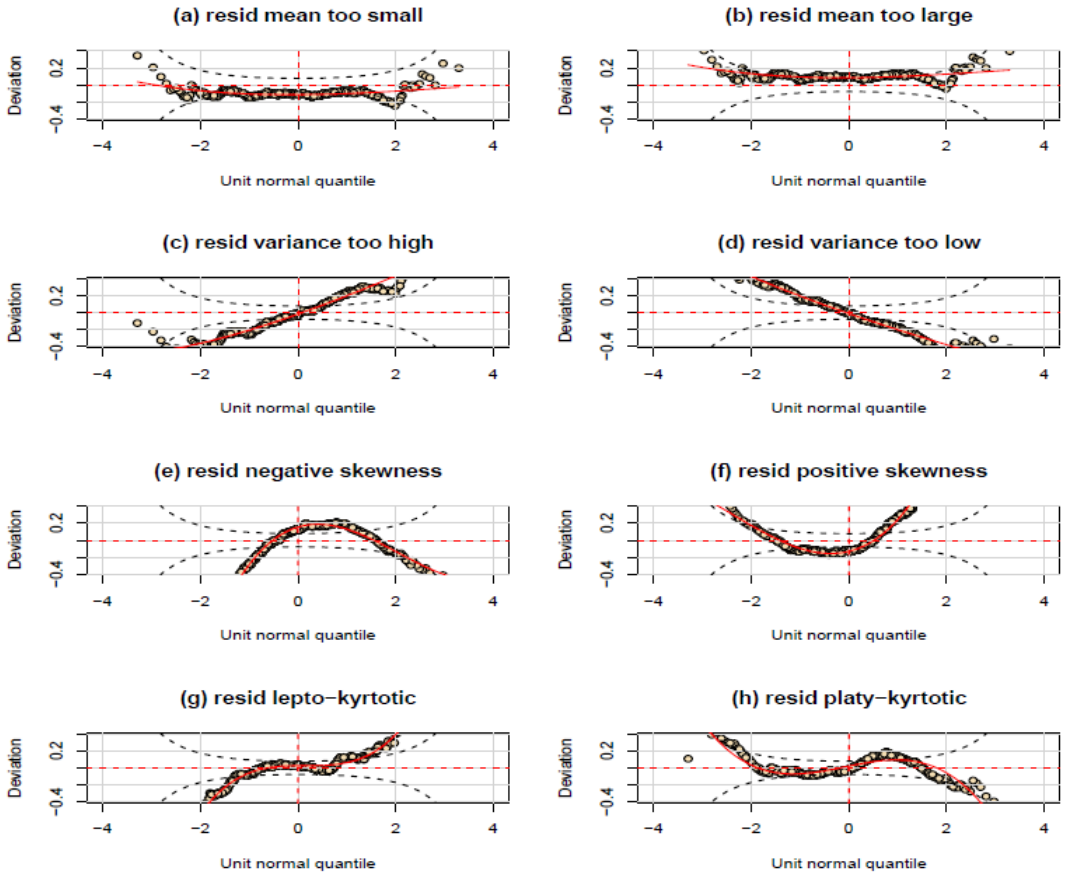

Figura 3.2: *Worm plot* de diferentes modelos GAMLSS incorretamente ajustados (Stasinopoulos et al., 2015).

3.6 Função gamlss()

Esta função é a principal do *package* GAMLSS, uma vez que é a responsável pelo ajustamento do modelo. Esta função deve obter os mesmos resultados que a função **glm**(), para modelos lineares generalizados, os mesmos valores ajustados e estimativas dos coeficientes para a modelação do valor médio. No entanto, o parâmetro de dispersão ϕ dos modelos **glm**() é estimado pelo *moment estimator*, enquanto pela função **gamlss**() o parâmetro de escala $\sigma = \phi^{\frac{1}{2}}$ é ajustado pela máxima verosimilhança (Stasinopoulos et al., 2015).

No caso de modelos com suavizadores, os resultados obtidos para o ajustamento do valor médio utilizando a função **gamlss**() deverão ser os mesmos resultados que os obtidos pela função **gam**() do *package* **gam**. No entanto, para tal acontecer, a função suavizadora utilizada pelo **gamlss**() deverá ser a **cs**(). Se a função utilizada for a **pb**(), as funções **gamlss**() e **gam**() deverão produzir resultados semelhantes utilizando o mesmo método de estimação do parâmetro de suavização. A função **gamlss**() utiliza o estimador de máxima verosimilhança local para o parâmetro suavizador, enquanto a função **gam**() utiliza validação-cruzada generalizada (Stasinopoulos et al., 2015).

Alguns dos argumentos e os respetivos valores por defeito da função **gamlss** são: formula = formula(data), sigma.formula = 1, nu.formula = 1, tau.formula = 1, family = NO(), method

= RS(), start.from = NULL, mu.start = NULL, sigma.start = NULL, nu.start = NULL, tau.start = NULL,...

A função **gamlss**() tem alguns argumentos que são necessários para a contrução do modelo. O conjunto dos termos inseridos para o parâmetro μ da distribuição, têm de ser representados no argumento *formula* da função. No entanto, para os restantes parâmetros da distribuição σ, ν e τ os argumentos utilizados são *sigma.formula*, *nu.formula* e *tau.formula*. Nesta função é necessário especificar a família da distribuição condicional da variável resposta pelo argumento *family*. O argumento *data* representa a *data.frame* dos dados utilizados para o estudo. É ainda possível especificar o algoritmo a ser utilizado para a estimação do modelo, **RS**(), **CG**() ou uma mistura de ambos, **mixed**(). Este último método inicia-se com o algoritmo de **RS**() e termina com o **CG**(). Para o caso de serem considerados alguns valores fixos para os parâmetros da distribuição estes podem ser definidos pelos argumentos *mu.fix*, *sigma.fix, nu.fix* e *tau.fix*, para os parâmetros μ σ, ν e τ. Em geral, os argumentos *mu.start*, *sigma.start*, *nu.start* e *tau.start* não têm necessidade de serem definidos. No entanto, para ajustar um novo modelo com valores de inicialização para os parâmetros da distribuição é possível defini-los através destes argumentos (Stasinopoulos et al., 2015).

3.7 Distribuições disponíveis no *package* GAMLSS

A distribuição assumida pela variável resposta Y pode ser muito geral, tendo uma única restrição imposta pelos modelos GAMLSS. A função $f(y|\boldsymbol{\theta})$ e a sua primeira derivada (opcionalmente, a segunda derivada e a derivada cruzada) têm de ser obtidas por métodos computacionais, em ordem a cada parâmetro θ_k da distribuição (Florencio, 2010).

O tipo da distribuição selecionada para a variável resposta depende da sua natureza. Existem três tipos de distribuições disponíveis: contínuas, discretas ou mistura de distribuições (Stasinopoulos et al., 2015). Uma vez que a variável resposta deste estudo é de natureza contínua, não serão abordadas as distribuições discretas e a mistura de distribuições. Para mais detalhes sobre esta informação consultar Stasinopoulos et al. (2015).

Na tabela 3.2, encontram-se exemplificadas algumas distribuições contínuas com um, dois, três ou quatro parâmetros da distribuição e respetivas funções de ligação, implementadas pelo *package* GAMLSS, no programa R (Rigby and Stasinopoulos, 2009).

As funções de ligação apresentadas são as utilizadas, por defeito, pela função **gamlss**() do *package* GAMLSS do programa R. Estas funções foram apresentadas por Wedderburn e Nelder, 1972, para os MLG de forma a garantir que o intervalo de valores para os parâmetros da distribuição se mantenham apropriados. Por exemplo, a distribuição Beta pode assumir valores definidos entre (0,1) e a função de ligação para o parâmetro μ é a função **logit**. O seu preditor será dado por $\eta = \log(\frac{\mu}{1-\mu})$, para ajustar o parâmetro μ pela seguinte forma $\mu = \dfrac{e^{\eta}}{1+e^{\eta}}$, que garante que a estimativa do μ seja entre 0 e 1 (Stasinopoulos et al., 2015).

Tabela 3.2: Algumas das distribuições contínuas disponíveis pelo *package* GAMLSS (Stasinopoulos et al., 2015)

		Função de ligação			
Distribuições	**Nomenclatura no R**	μ	σ	ν	τ
beta	BE()	logit	logit	-	-
beta inflacionada (em 0)	BEOI()	logit	log	logit	-
beta inflacionada (em 1)	BEZI()	logit	log	logit	-
beta inflacionada (em 0 e 1)	BEINF()	logit	logit	log	log
Box-Cox (Cole & Green)	BCCG()	identidade	log	identidade	-
Box-Cox exponencial potência	BCPE()	identidade	log	identidade	log
Box-Cox-t	BCT()	identidade	log	identidade	log
exponencial	EXP()	log	-	-	-
exponencial gaussiana	exGAUS()	identidade	log	log	-
exponencial power	PE()	identidade	log	log	-
gama	GA()	log	log	-	-
gama generalizada	GG()	log	log	identidade	-
beta generalizada tipo 1	GB1()	logit	logit	log	log
beta generalizada tipo 2	GB2()	log	identidade	log	log
gaussiana inversa	IG()	log	log	-	-
gama inversa	IGAMMA()	log	log	-	-
gaussiana inversa generalizada	GIG()	log	log	identidade	-
Gumbel	GU()	identidade	log	-	-
log normal	LOGNO()	log	log	-	-
log normal (Box-Cox)	LNO()	log	log	fixa	-
logística	LO()	identidade	log	-	-
normal	NO()	identidade	log	-	-
Pareto 2	PARETO2()	log	log	-	-
Weibull	WEI()	log	log	-	-
Weibull (μ a média)	WEI3()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-

As distribuições contínuas podem ser simétricas, assimétricas positivas/negativas e em relação à curtose, mesocúrticas (*mesokurtic*), leptocúrticas (*leptokurtic*) ou platicúrticas (*platykurtic*). Na figura 3.3 é possível observar quatro gráficos onde estão representadas diferentes formas das distribuições: assimétrica positiva, assimétrica negativa, *platykurtic* e *leptokurtic*. A análise da curtose é feita em comparação com uma distribuição normal (0,1), com forma *mesokurtic* (Stasinopoulos and Rigby, 2014).

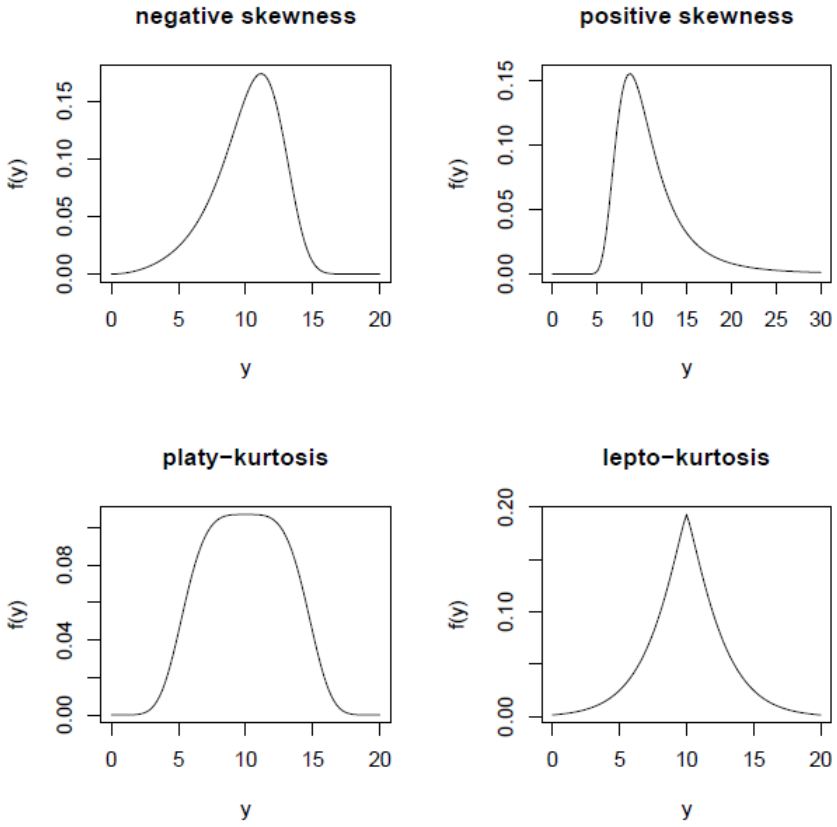

Figura 3.3: Diferentes tipos de distribuições contínuas: assimetria negativa (gráfico do canto superior esquerdo), assimetria positiva (gráfico do canto superior direito), *platykurtic* (gráfico do canto inferior esquerdo) e *leptokurtic* (gráfico do canto inferior direito) (Stasinopoulos and Rigby, 2014).

O suporte das distribuições contínuas no *package* GAMLSS divide as distribuições em três grupos: (-∞, +∞), (0,+∞) e (0,1) (Rigby and Stasinopoulos, 2009). Como exemplo, serão demonstradas duas distribuições da família da distribuição beta generalizada, tipo 1 e tipo 2. A primeira tem um suporte (0,1), enquanto a segunda tem suporte (0,+∞) (Rigby et al., 2014).

- **Distribuição Generalizada Beta tipo 1 (GB1)**

A família da distribuição beta generalizada tipo 1, tem suporte $0 < Y < 1$, GB1(μ, σ, ν, τ), e define uma variável Z pertencente a uma distribuição Beta, BE(μ, σ) (Rigby et al., 2014):

$$Z = \frac{Y^{\tau}}{\nu + (1-\nu)Y^{\tau}}, \tag{3.22}$$

onde $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ e $\tau > 0$ (Rigby et al., 2014). A função densidade de GB1

é dada por:

$$f(y|\mu,\sigma,\nu,\tau) = \frac{\tau\nu^{\beta}y^{\tau\alpha-1}(1-y^{\tau})^{\beta-1}}{B(\alpha,\beta)[\nu+(1-\nu)y^{\tau}]^{\alpha+\beta}}, \tag{3.23}$$

em que $\alpha = \mu(1-\sigma^2)/\sigma^2$ e $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $\alpha > 0$ e $\beta > 0$. Os parâmetros μ e σ são adaptados para $\mu = \alpha/(\alpha+\beta)$ e $\sigma = (\alpha+\beta+1)^{-1/2}$. A função BE$(\mu,\sigma)$ é um caso especial de GB1(μ,σ,ν,τ), em que $\nu = 1$ e $\tau = 1$ (Rigby et al., 2014).

- **Distribuição Generalizada Beta tipo 2 (GB2)**

Quanto à função tipo 2, a função densidade GB2 (μ,σ,ν,τ) é dada por:

$$f(y|\mu,\sigma,\nu,\tau) = |\sigma|y^{\sigma\nu-1}\{\mu^{\sigma\nu}B(\nu,\tau)[1+(y/\mu)^{\sigma}]^{\nu+\tau}\}^{-1} \tag{3.24}$$

$$= \frac{\Gamma(\nu+\tau)}{\Gamma(\nu)\Gamma(\tau)}\frac{\sigma(y/\mu)^{\sigma\nu}}{y[1+(y/\mu)^{\sigma}]^{\nu+\tau}}, \tag{3.25}$$

onde $y > 0$, $\mu > 0$, $-\infty < \sigma < \infty$, $\nu > 0$ e $\tau > 0$. O valor médio e a variância de Y são calculados a partir de $E(Y) = \mu B(\nu+\frac{1}{\sigma},\tau-\frac{1}{\sigma})/B(\nu,\tau)$, para $-\nu < \frac{1}{\sigma} < \tau$ e de $E(Y^2) = \mu^2 B(\nu+\frac{2}{\sigma},\tau-\frac{2}{\sigma})/B(\nu,\tau)$ para $-\nu < \frac{2}{\sigma} < \tau$ (Rigby et al., 2014).

Para o caso de $\nu = 1$ é obtida a distribuição de Burr:

$$f(y|\mu,\sigma,\nu,\tau) = \frac{\tau\sigma(y/\mu)^{\sigma}}{y[1+(y/\mu)^{\sigma}]^{\tau+1}}. \tag{3.26}$$

Se $\sigma = 1$ é obtida a distribuição Pareto Generalizada:

$$f(y|\mu,\sigma,\nu,\tau) = \frac{\Gamma(\nu+\tau)}{\Gamma(\nu)+\Gamma(\tau)}\frac{\mu^{\tau}y^{\nu-1}}{(y+\mu)^{\nu+\tau}}. \tag{3.27}$$

Capítulo 4

Análise dos Dados

4.1 Amostra

Os dados recolhidos e utilizados para este estudo foram previamente analisados e os resultados, publicados em 2010 (Soto et al., 2010) e em 2013 (Soto et al., 2013). Nenhum estudo foi feito no âmbito da estimação da creatinina sérica basal.

O grupo de investigação, a que pertence a Doutora Karina Soto, nefrologista no HFF, investigadora principal, e Professora Doutora Ana Luísa Papoila, Bioestatista na NOVA Medical School/Faculdade de Ciências Médicas da Universidade Nova de Lisboa, responsável pela análise estatística dos dados, foi o responsável pela recolha da informação utilizada neste estudo. Para um doente ser incluído no estudo deveria ter sido admitido na unidade de emergência médica não cirúrgica do HFF entre março e novembro de 2008, após a obtenção do consentimento informado. Em relação aos critérios de exclusão, não foram incluídos no estudo todos os doentes que verificam o seguinte: (1) menores de 18 anos ou maiores de 80 anos; (2) com doença renal crónica em estádio 4 (CKD-4); (3) com anúria (perda total da produção de urina) completa; (4) com obstrução urinária; (5) em tratamento de quimioterapia; (6) com AKI em estádio 3 classificado pelo critério AKIN; (7) com internamento inferior a 48h (Soto et al., 2010).

4.2 *Software*

Toda a análise dos dados foi implementada no programa estatístico R, versão 3.3.1. As bibliotecas utilizadas foram: **gamlss**, **AID** e **gam**. Em apêndice, encontra-se o *script* de toda a análise efetuada neste estudo.

4.3 Variáveis

Foram incluídos no estudo 600 indivíduos e recolhidos os valores de 124 variáveis, entre as quais a idade, o sexo, a raça, a creatinina sérica basal e a creatinina sérica à data de admissão. O valor da creatinina sérica basal foi obtido através da informação contida nos processos clínicos dos doentes, seguidos em consulta, entre 1 a 6 meses anteriores à data de admissão na unidade hospitalar (Soto et al., 2010).

Devido ao elevado número das variáveis consideradas para a realização dos estudos já referidos, a base de dados utilizada passou a conter apenas as variáveis de interesse para este estudo. Desta forma foi criada uma nova base de dados mais simplificada. Não foram obtidos valores omissos para qualquer uma das variáveis que passamos a apresentar:

CREATININA SÉRICA BASAL (scrb) - variável quantitativa contínua, medida na unidade mg/dl. Indica o valor da creatinina sérica basal do indivíduo, num período de 1 a 6 meses anteriores à data da inclusão no estudo.

CREATININA SÉRICA NO TEMPO 0 (scr_t0) - variável quantitativa contínua, medida na unidade mg/dl. Indica o valor da creatinina sérica à data de admissão do indivíduo na unidade hospitalar.

SEXO (sexo) - variável qualitativa binária que representa o sexo do indivíduo.

IDADE (idade) - variável quantitativa contínua que representa a idade do indivíduo, em anos, à data de admissão na unidade hospitalar Fernando Fonseca.

RAÇA (raca) - variável qualitativa binária que representa a raça do indivíduo.

Na tabela 4.1 encontram-se as variáveis utilizadas neste estudo. A coluna **Nome** corresponde ao nome de codificação de cada uma no *script* do R.

Tabela 4.1: Classificação e codificação das variáveis

Nome	**Descrição**	**Tipo**	**Codificação**
scrb	Creatinina sérica basal	Quantitativa contínua	
scr_t0	Creatinina sérica no tempo 0	Quantitativa contínua	
sexo	Sexo	Qualitativa Nominal	1-Masculino 2-Feminino
raca	Raça	Qualitativa Nominal	1-não Negra 2-Negra
idade	Idade	Quantitativa contínua	

4.4 Análise Exploratória

4.4.1 Variável Dependente

As medidas descritivas para a variável dependente, creatinina sérica basal, encontram-se na tabela 4.2.

Tabela 4.2: Medidas descritivas da creatinina sérica basal

Mínimo	1º quartil	Média	Mediana	3º quartil	Máximo	Desvio padrão
0.3	0.6	0.8197	0.8	0.9	2.7	0.2993

Para esta variável ainda foi construídos um diagrama em caixa de bigodes e um histograma (Figura 4.1). Ambos sugerem que a creatinina sérica basal não segue uma distribuição normal, uma vez que é observado uma assimetria à direita. Além disso, surgem ainda 11 valores extremos (Figura 4.1 - direita), facto este favorável à premissa de que estes dados não provêm de uma população com distribuição normal.

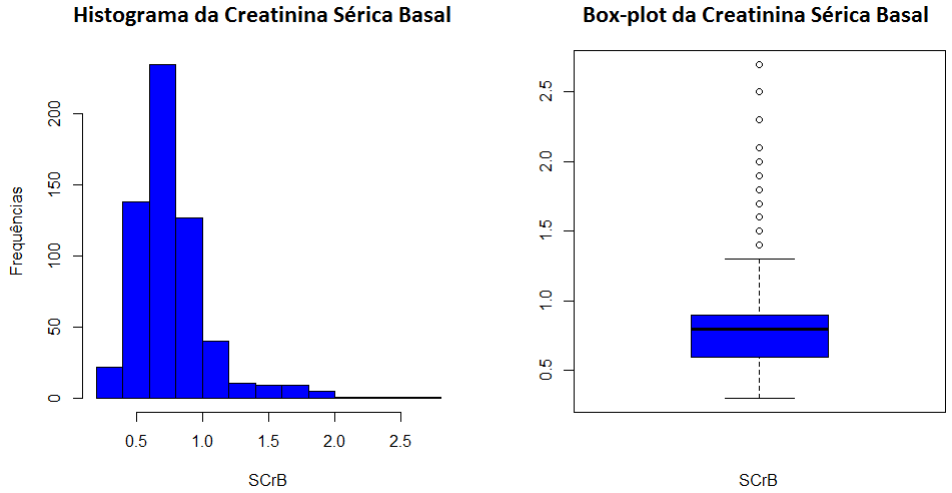

Figura 4.1: Histograma (esquerda) e diagrama em caixa de bigodes (direita) da variável creatinina sérica basal.

Para esta variável foi realizada uma transformação logarítmica na tentativa de obter normalidade da sua distribuição. Pela visualização do histograma da variável (Figura 4.2 - esquerda), parece observar-se alguma simetria na distribuição dos dados, no entanto, o diagrama em caixa de bigodes (Figura 4.2 - direita), evidência uma certa assimetria, e também a presença de diversas observações extremas.

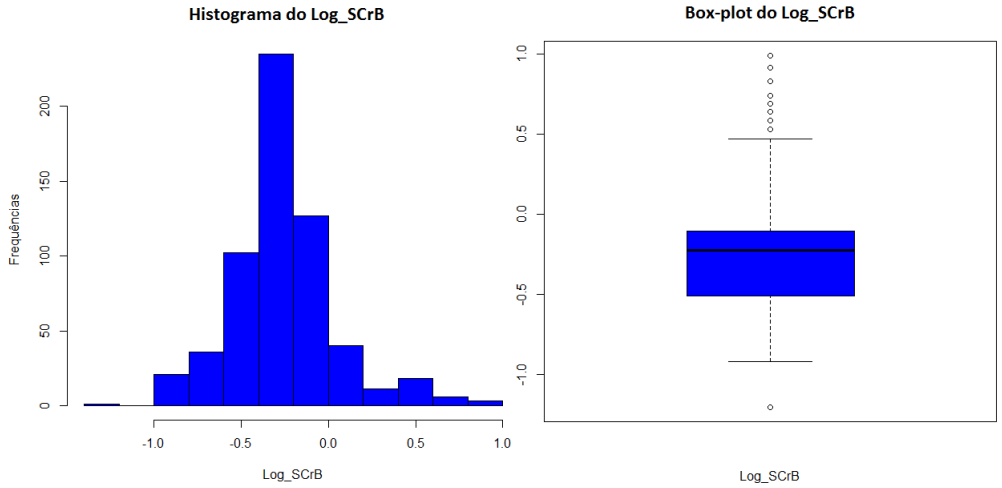

Figura 4.2: Histograma (esquerda) e diagrama em caixa de bigodes (direita) da variável logaritmo da creatinina sérica basal.

Na tabela 4.3 encontram-se algumas medidas descritivas da transformação logarítmica da variável creatinina sérica basal.

Tabela 4.3: Medidas descritivas do logaritmo da creatinina sérica basal

Mínimo	**1º quartil**	**Média**	**Mediana**	**3º quartil**	**Máximo**	**Desvio padrão**
-1.2040	-0.5108	-0.2519	-0.2232	-0.1054	0.9933	0.3146

4.4.2 Variáveis Independentes

4.4.2.1 Sexo, Raça

Em relação às variáveis sexo e raça é possível observar pela tabela 4.4, que os indivíduos da amostra não se distribuem de forma equitativa entre as categorias destas variáveis independentes. As diferenças das proporções entre as raças são mais significativas (87% e 13%) do que entre as categorias do sexo (38.33% e 61.67%).

Tabela 4.4: Proporções de indivíduos para as categorias das variáveis sexo e raça

Sexo	**Total (n=600)**	Raça	**Total (n=600)**
Feminino (%)	230 (38.33%)	**não Negra (%)**	522 (87%)
Masculino (%)	370 (61.67%)	**Negra (%)**	78 (13%)

Foram ainda construídos quatro diagramas em caixa de bigodes para os valores da creatinina sérica basal (Figura 4.3), das categorias do sexo feminino e masculino e das raças não negra e negra. Em todos os diagramas em caixa de bigodes é possível observar diversos valores extremos.

Nos diagramas em caixa de bigodes da variável sexo é possível observar uma simetria em ambos os gráficos para a categoria do sexo feminino e do sexo masculino. No entanto os valores

da SCr basal do sexo masculino são ligeiramente superiores aos do sexo feminino sugerindo uma provável diferença entre os valores médios da SCr dos dois sexos.

O diagrama em caixa de bigodes da raça negra também apresenta uma simetria dos valores da creatinina sérica basal, no entanto o mesmo não ocorre na raça não negra. A raça negra é a que apresenta menos valores extremos e uma simetria dos valores da creatinina sérica basal.

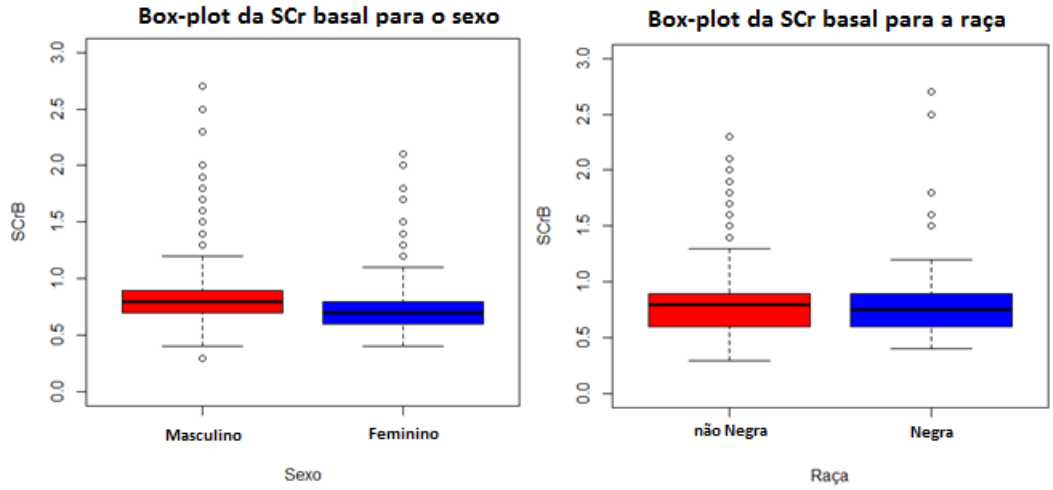

Figura 4.3: Diagrama em caixa de bigodes da creatinina sérica basal para as categorias das variáveis sexo (feminino e masculino) e raça (não negra e negra).

Na tabela tabela 4.5 encontram-se algumas medidas descritivas da creatinina sérica basal para as diferentes categorias das variáveis sexo e raça.

Tabela 4.5: Medidas descritivas da creatinina sérica basal para as categorias das variáveis sexo e raça

Categorias	**Mín.**	**1º quartil**	**Média**	**Mediana**	**3º quartil**	**Máx.**	**Desv. padrão**
Fem.	0.4	0.6	0.7557	0.7	0.8	2.1	0.2858
Masc.	0.3	0.7	0.8595	0.8	0.9	2.7	0.3009
não-Negra	0.3	0.6	0.8188	0.8	0.9	2.3	0.2839
Negra	0.4	0.6	0.8256	0.75	0.875	2.7	0.3889

4.4.2.2 Idade

Na tabela 4.6 encontram-se algumas medidas descritivas desta variável, onde é possível observar que os pacientes incluídos no estudo têm uma idade mediana de 62 anos. O histograma e o diagrama em caixa de bigodes (Figura 4.4) evidenciam que esta variável provavelmente não tem uma distribuição normal, uma vez que se nota uma maior assimetria à esquerda.

Tabela 4.6: Medidas descritivas da idade

Mínimo	1º quartil	Média	Mediana	3º quartil	Máximo	Desvio padrão
18	50	59.1	62	71	81	15.51

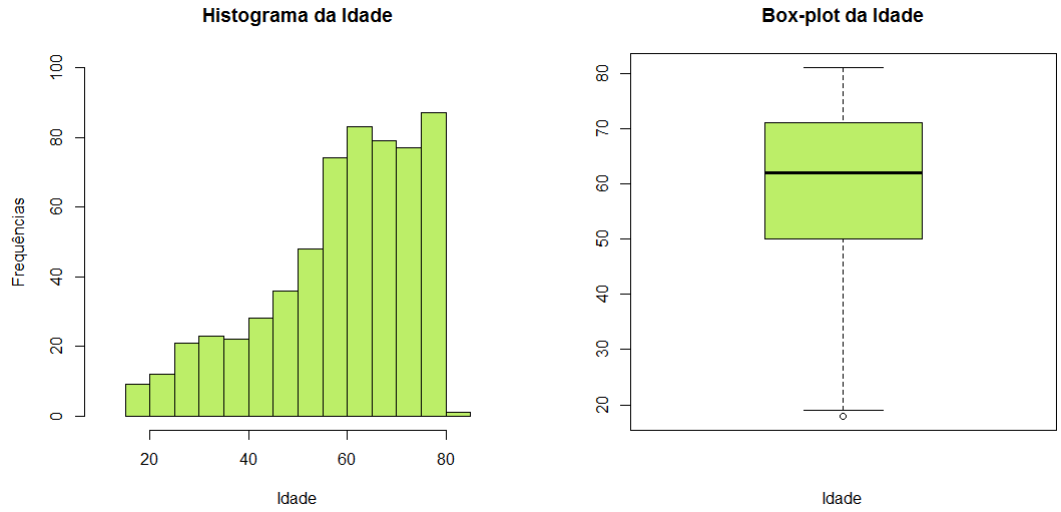

Figura 4.4: Histograma (esquerda) e diagrama em caixa de bigodes (direita) da variável idade.

4.4.2.3 Creatinina Sérica no tempo 0

Na tabela 4.7 encontram-se algumas medidas descritivas da variável creatinina sérica no tempo 0. Como é possível observar, os valores desta variável têm uma amplitude superior aos valores da variável creatinina sérica basal, a variável dependente.

Tabela 4.7: Medidas descritivas da creatinina sérica no tempo 0

Mínimo	1º quartil	Média	Mediana	3º quartil	Máximo	Des. padrão
0.2	0.7	1.109	0.9	1.3	6.4	0.7024

O diagrama em caixa de bigodes e o histograma (Figura 4.5) mostram uma assimetria à direita, sugerindo que esta variável não tem uma distribuição normal. Ainda é possível observar um elevado número de valores extremos superior a 2 mg/dL.

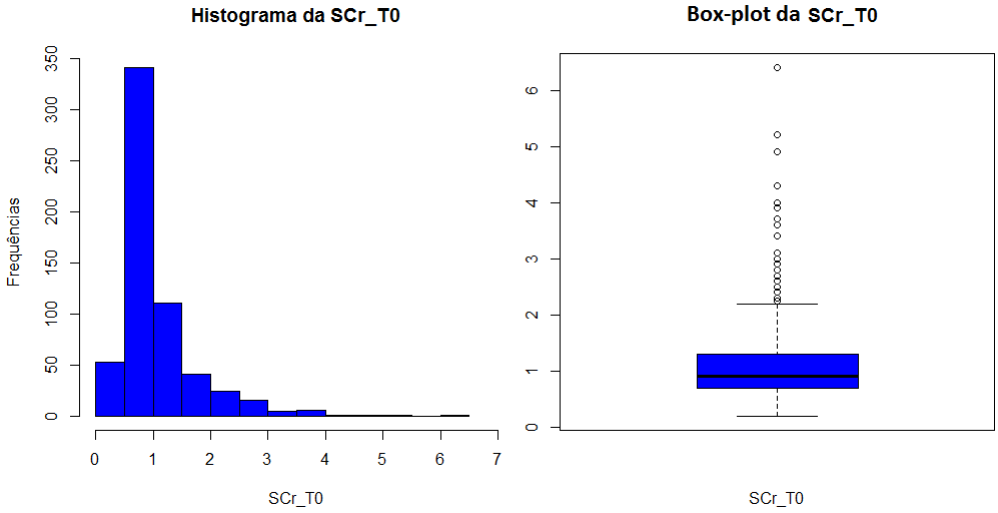

Figura 4.5: Histograma (esquerda) e diagrama em caixa de bigodes (direita) da variável creatinina sérica no tempo 0.

4.5 Inferência Estatística

Para analisar se as amostras, relativas às variáveis contínuas, provêm de uma população com uma distribuição normal, foi realizado o teste de ajustamento de Kolmogorov-Smirnov. Este teste foi escolhido uma vez que o tamanho da amostra é elevado. Foram testadas as distribuições das variáveis idade, creatinina sérica no tempo 0 e creatinina sérica basal. A hipótese nula, caso não seja rejeitada, permite-nos concluir que as populações de onde foi retirada a amostra têm uma distribuição normal. Para além destas variáveis ainda foi avaliada a normalidade no caso da transformação logarítmica da creatinina sérica basal e das categorias da variável sexo e raça.

Na tabela 4.8 encontram-se todos os valores da estatística de teste e os valores-p de todos os testes aplicados. Como é possível observar, a normalidade da distribuição das variáveis foi rejeitada, para todas as variáveis. Embora os diagramas em caixa de bigodes da SCr basal para as categorias das variáveis sexo e raça sugerissem uma provável normalidade da distribuição, esta conclusão foi descartada devido aos resultados do teste de ajustamento.

Tabela 4.8: Teste de Kolmogorov-Smirnov

	Teste de Kolmogorov-Smirnov	
Variável	**Estatística de Teste**	**Valor-p**
SCr Basal	0.19418	2.2×10^{-16}
SCr no tempo 0	0.21817	2.2×10^{-16}
Idade	0.097342	2.306×10^{-5}
Logaritmo da SCr Basal	0.12193	3.569×10^{-8}
SCr Basal sexo Feminino	0.22096	3.529×10^{-10}
SCr Basal sexo Masculino	0.21939	6.661×10^{-16}
SCr Basal raça não Negra	0.17859	6.883×10^{-16}
SCr Basal raça Negra	0.28317	7.389×10^{-6}

Foram, ainda, construídos gráficos *QQ-plot* (Figura 4.6) para as variáveis contínuas (idade, creatinina no tempo zero e creatinina basal e respetiva transformação logarítmica) de forma a verificar o ajustamento dos dados à distribuição normal. Em todos os gráficos verificou-se que todas as variáveis contínuas apresentam uma fuga à normalidade da sua distribuição.

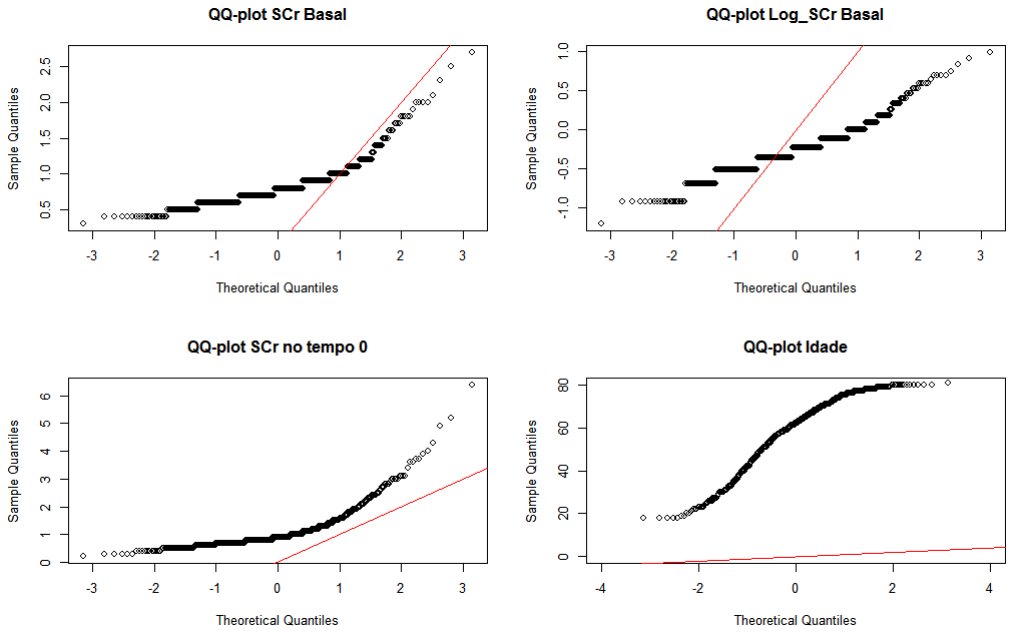

Figura 4.6: *QQ-plot* para as variáveis creatinina sérica basal (canto superior esquerdo), transformação logarítmica da creatinina sérica basal (canto superior direito), creatinina sérica no tempo 0 (canto inferior esquerdo) e idade (canto inferior direito).

Os diferentes *QQ-plots* construídos (Figura 4.7) para os sexos feminino e masculino e para as raças negra e não negra, sugerem que a creatinina sérica basal para estas categorias não apresenta uma distribuição normal. Embora os diagramas em caixa de bigodes, apresentados anteriormente, sugerissem uma possível normalidade, o teste de Kolmogorov-Smirnov e os *QQ-plots* vieram comprovar o contrário. Provavelmente, o grande número de valores extremos dificulta a possibilidade desta variável ter uma distribuição normal.

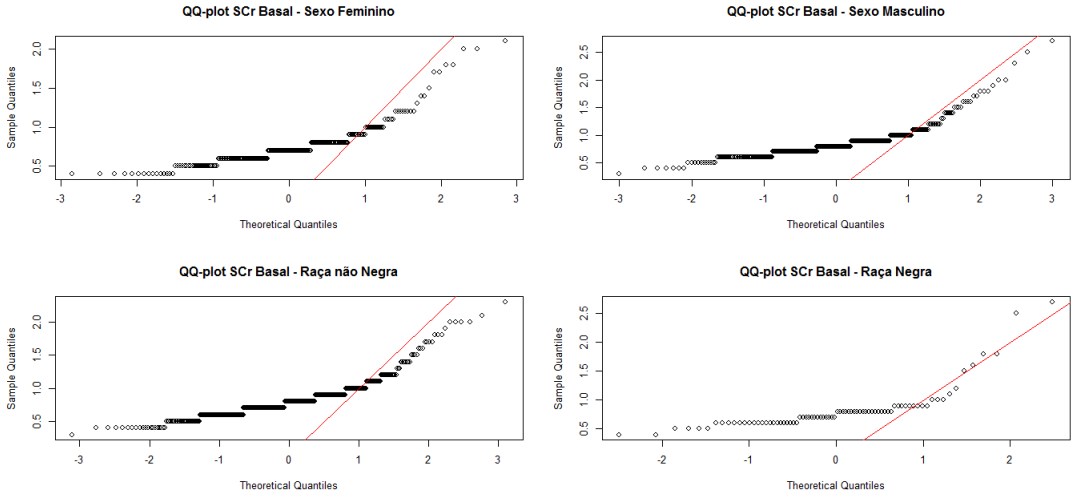

Figura 4.7: *QQ-plot* para a variável creatinina sérica basal das categorias do sexo (feminino - canto superior esquerdo; masculino - canto superior direito), da raça não negra (canto inferior esquerdo) e da raça negra (canto inferior direito).

Ainda foi testado se haveria diferenças entre os valores médios da creatinina sérica basal, entre as categorias das variáveis sexo e raça. O teste utilizado foi o teste-Z, obtendo-se os seguintes resultados apresentados na tabela 4.9. Concluiu-se que existem diferenças significativas entre os valores médios da creatinina sérica basal do sexo feminino em relação à do sexo masculino. No entanto, para a variável raça não é possível afirmar que existam diferenças significativas entre os valores médios das raça não negra e a raça negra.

Tabela 4.9: Teste-t para a diferença entre os valores médios da creatinina sérica basal entre as categorias das variáveis sexo e raça

	Sexo	**Raça**
Estatística de teste-Z	4.1876	-0.1501
Valor-p	3.243×10^{-5}	0.8810

Capítulo 5

Análise dos Dados através de um MLG e de um MAG

O objetivo deste estudo é obter um modelo de predição para o valor da creatinina sérica basal. Assim, iniciou-se a análise utilizando-se um modelo linear generalizado e, posteriormente, um modelo aditivo generalizado. Para ambos, foram inseridas todas as variáveis independentes, idade, sexo, raça e creatinina no tempo 0, uma vez que todas têm um significado clínico na área da lesão renal aguda.

5.1 Modelo Linear Generalizado

O modelo linear generalizado obtido tem a seguinte expressão:

$$g(\mu) = \eta = \beta_0 + \beta_1 \text{idade} + \beta_2 \text{raca} + \beta_3 \text{sexo} + \beta_4 \text{scr\_}t0. \tag{5.1}$$

Para avaliar a qualidade do ajustamento do modelo, foram analisados os resíduos através de um *QQ-plot*. O gráfico obtido sugere que a distribuição dos resíduos não é uma distribuição normal (Figura 5.1). Uma vez que esta condição de aplicabilidade do modelo não se verifica iremos seguir uma nova abordagem, após efetuar uma transformação da variável resposta.

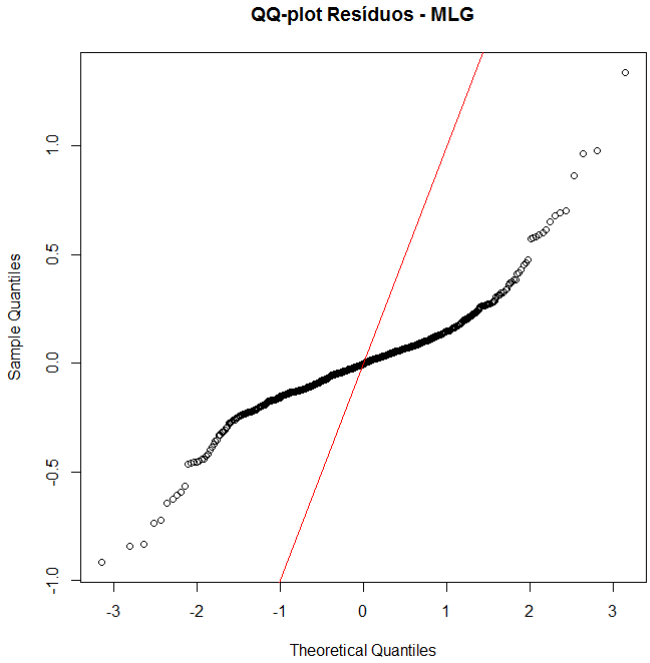

Figura 5.1: *QQ-plot* dos resíduos obtidos com o modelo de regressão linear.

Assim sendo, procedeu-se à transformação logarítmica da variável resposta. A expressão do modelo obtido é a seguinte:

$$g(log(\mu)) = \eta = \beta_0 + \beta_1 \text{idade} + \beta_2 \text{raca} + \beta_3 \text{sexo} + \beta_4 \text{scr\_}t0. \tag{5.2}$$

O gráfico *QQ-plot* obtido (Figura 5.2) mostra uma melhoria da distribuição dos resíduos. No entanto, como no modelo anterior, não é possível concluir que os resíduos sejam provenientes de uma população com distribuição normal.

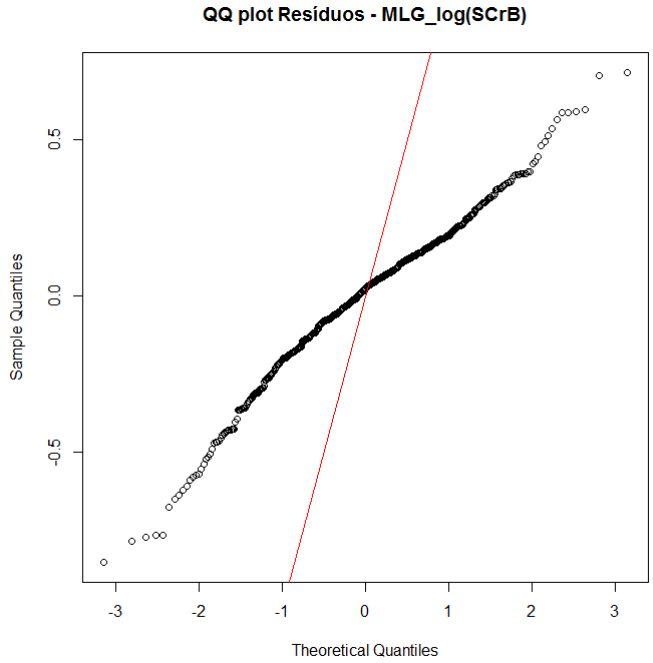

Figura 5.2: *QQ-plot* dos resíduos obtidos pelo modelo de regressão linear com transformação logarítmica da variável dependente, creatinina sérica basal.

Uma vez que a condição de aplicabilidade respeitante à normalidade dos resíduos não se verificou, foi necessário recorrer a outra estratégia utilizando a transformação de Box-Cox. Através desta transformação da variável dependente é necessário estimar o parâmetro λ que, em algumas circunstâncias, permite normalizar a distribuição daquela variável. Através dos resultados obtidos pela função **boxcoxnc**() (Tabela 5.1), é possível verificar que, para todos os testes de normalidade realizados, foi rejeitada a hipótese de normalidade da distribuição da SCr basal, para qualquer parâmetro de transformação Box-Cox estimado, $\hat{\lambda}$. A primeira linha da tabela indica a estimativa de λ, as linhas seguintes indicam os valores-p obtidos com os três diferentes testes de normalidade, *sw* corresponde ao teste de Shapiro-Wilk, *sf* ao teste de Shapiro-Francia e *jb* ao teste de Jarque-Bera. Os diferentes métodos de estimação do parâmetro λ estão identificados ao topo de cada coluna, **SW** corresponde ao método de Shapiro-Wilk, **AD** ao método de Anderson-Darling, **CVM** ao método de Cramer-von Mises, **PT** Pearson Chi-square, **SF** Shapiro-Francia, **LT** Lilliefords, **JB** Jarque-Bera e **AC** ao método da covariável artificial.

Tabela 5.1: Transformação de Box-Cox. **SW**: Shapiro-Wilk; **AD**:Anderson-Darling; **CVM**: Cramer-von Mises; **PT**: Pearson Chi-square; **SF**: Shapiro-Francia; **LT**: Lilliefords; **JB**: Jarque-Bera; **AC** - método da covariável artificial

	SW	**AD**	**CVM**	**PT**
lambda.hat	-4.300000e-01	-4.400000e-01	-4.500000e-01	-1.580000e+00
sw.pvalue	2.047634e-10	2.034419e-10	2.007389e-10	6.182567e-21
sf.pvalue	1.754947e-09	1.744717e-09	1.724244e-09	1.593638e-18
jb.pvalue	1.064312e-04	1.065391e-04	1.036870e-04	0.000000e+00
	SF	**LT**	**JB**	**AC**
lambda.hat	-4.300000e-01	-1.800000e-01	-4.400000e-01	-4.603333e-01
sw.pvalue	2.047634e-10	2.650608e-11	2.034419e-10	1.965598e-10
sf.pvalue	1.754947e-09	3.025959e-10	1.744717e-09	1.692764e-09
jb.pvalue	1.064312e-04	5.559996e-09	1.065391e-04	9.789084e-05

5.2 Modelo Aditivo Generalizado

Uma vez que não se obteve a condição de normalidade da distribuição dos resíduos para ambos os modelos lineares generalizados e não foi possível estimar o parâmetro da transformação de Box-Cox, recorreu-se aos modelos aditivos generalizados.

Inicialmente foram construídos os gráficos das funções parciais dos modelos univariados, entre a creatinina sérica basal e as variáveis contínuas independentes, idade e creatinina sérica no tempo 0. Através do gráfico da figura 5.3 - esquerda, a relação entre a variável idade e a creatinina sérica basal é praticamente linear, sugerindo a não utilização de um suavizador para esta variável independente. No entanto, no gráfico da creatinina sérica no tempo 0 (Figura 5.3 - direita), verifica-se que não existe uma relação linear, sendo aconselhada a utilização de um suavizador para esta variável independente.

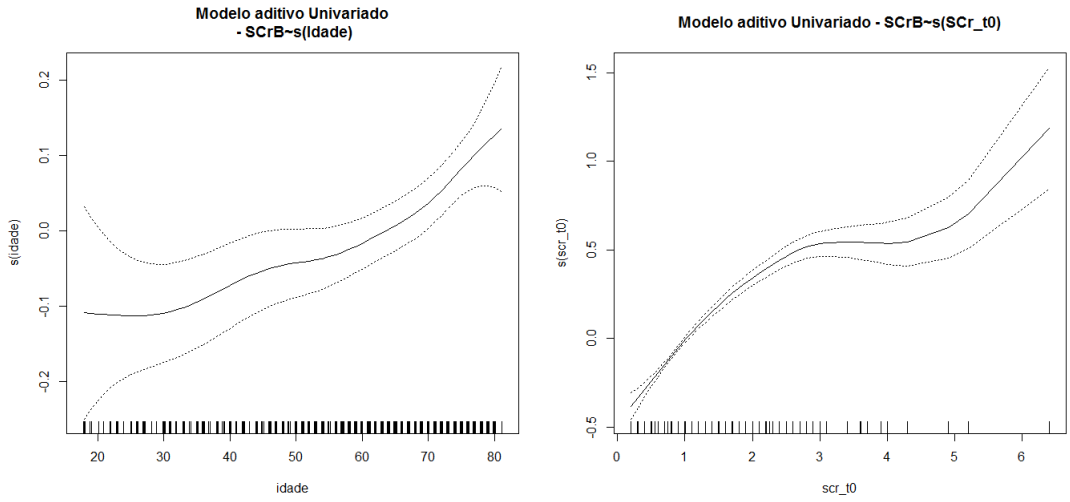

Figura 5.3: Gráficos das funções parciais da idade (esquerda) e da creatinina sérica no tempo 0 (direita) obtidos na análise univariada.

Assim, o modelo aditivo generalizado construído terá a seguinte expressão:

$$g(\mu) = \beta_1 \text{raca} + \beta_2 \text{sexo} + \beta_3 \text{idade} + \text{s}(\text{scr\_}t0) \tag{5.3}$$

Os resíduos obtidos pelo modelo aditivo generalizado, apresentados num gráfico *QQ-plot*, (Figura 5.4), não apresentam uma distribuição normal. Uma vez que a condição de aplicabilidade do modelo, em relação à normalidade dos resíduos falhou, a análise dos dados através deste modelo deu-se como terminada. A abordagem seguinte, para tentar ultrapassar a dificuldade de atingir a normalidade dos resíduos, constam num ajustamento de um modelo GAMLSS, cujo os resultados serão apresentados no capítulo seguinte.

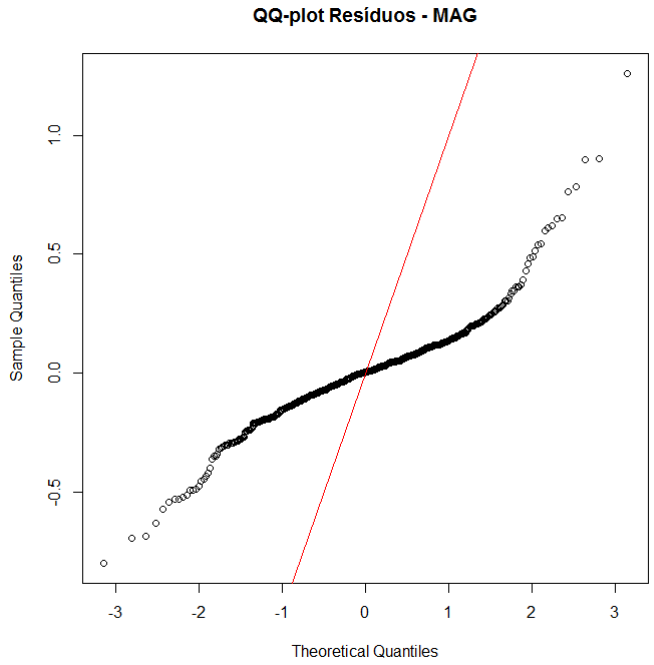

Figura 5.4: *QQ-plot* dos resíduos obtidos pelo modelo aditivo generalizado.

Capítulo 6

Análise dos Dados através de um GAMLSS

6.1 Escolha da distribuição da variável dependente

Após a não obtenção da normalidade dos resíduos dos modelos MLG e MAG, analisaram-se os dados através de modelos GAMLSS. Uma vez que estes modelos têm uma maior flexibilidade no que diz respeito à distribuição da variável resposta acreditamos que a falha do pressuposto acerca dos resíduos obtidos nos modelos anteriores possa ser ultrapassada.

A primeira análise utilizando o *package* GAMLSS constou da estimação da distribuição da variável creatinina sérica basal. A função **fitDist**() contém o argumento *type*, que teve necessidade de ser especificado para ajustar as funções contínuas com suporte em $\mathbb{R}^+$ à variável dependente. A opção '*realplus*' selecionada não contém todas as distribuições contínuas com suporte $\mathbb{R}^+$ disponíveis no *package* GAMLSS, desta forma foi necessário adicionar as restantes distribuições no argumento *extra*. Assim sendo, foram considerados as diferentes distribuições: *GB2*, *BCT*, *BCTo*, *exGAUS*, *BCPE*, *BCPEo*, *BCCG*, *BCCGo*, *IGAMMA*, *GG*, *GIG*, *LOGNO*, *IG*, *GA*, *WEI2*, *WEI*, *WEI3*, *EXP*, *PARETO2*. As distribuições *BCCGo*, *BCPEo* e *BCTo* têm a função de ligação logarítmica por *default*. Estas distribuições estão apresentadas na tabela 3.2.

Como é possível visualizar no *script* da figura 6.1, a distribuição obtida com menor valor de GAIC foi a *GB2 - generalized beta type 2*. Todas as distribuições testadas conseguiram ser ajustadas à variável dependente (sem problemas computacionais), obtendo-se sempre um valor de GAIC para cada uma das distribuições. A distribuição *GB2 - generalized beta type 2* é uma distribuição que contém quatro parâmetros, μ, σ, ν e τ, que serão modelados no ajustamento do modelo GAMLSS.

Após selecionada a distribuição da variável dependente, foi construído um histograma para os valores da creatinina sérica basal através da função **histDist**() (Figura 6.2). No histograma estão representados os valores observados da creatinina sérica basal, a linha a vermelho representa a função densidade paramétrica *GB2* ajustada à SCr basal, e a linha azul a densidade estimada não-parametricamente (através de um suavizador de Kernel). A função densidade *GB2* utiliza estimativas para ajustar a distribuição aos dados, tendo sido obtidos os valores $\hat{\mu} = -0.3462$,

$\hat{\sigma} = 9.935$, $\hat{\nu} = -0.4253$ e $\hat{\tau} = -0.8115$. As funções de ligação da *GB2* são, para μ, ν e τ a logarítmica e para σ a identidade.

```
> fitting_dist<-fitDist(scrb,type=c('realplus'),
+ extra=c('BCPE','BCCG','BCT','GB2', 'WEI','WEI2'))
Warning messages:
1: In MLE(ll2, start = list(eta.mu = eta.mu, eta.sigma = eta.sigma),  :
  possible convergence problem: optim gave code=1 false convergence (8)
2: In MLE(ll4, start = list(eta.mu = eta.mu, eta.sigma = eta.sigma,  :
  possible convergence problem: optim gave code=1 false convergence (8)
>
> fitting_dist$fits
       GB2      BCPEo       BCTo        BCT     exGAUS       BCPE      BCCGo
-28.583098 -25.484492 -25.012617 -25.012617 -16.047705 -12.396234  -7.716246
      BCCG     IGAMMA         GG        GIG      LOGNO         IG         GA
 -7.716246  -6.703682  -5.999659  -4.703682  15.641111  20.640133  68.701719
      WEI2       WEI       WEI3       EXP    PARETO2
247.693775 247.693775 247.693775 963.370969 965.370995
> fitting_dist$failed
list()
```

Figura 6.1: *Script* e respetivo *output* da função **fitDist**().

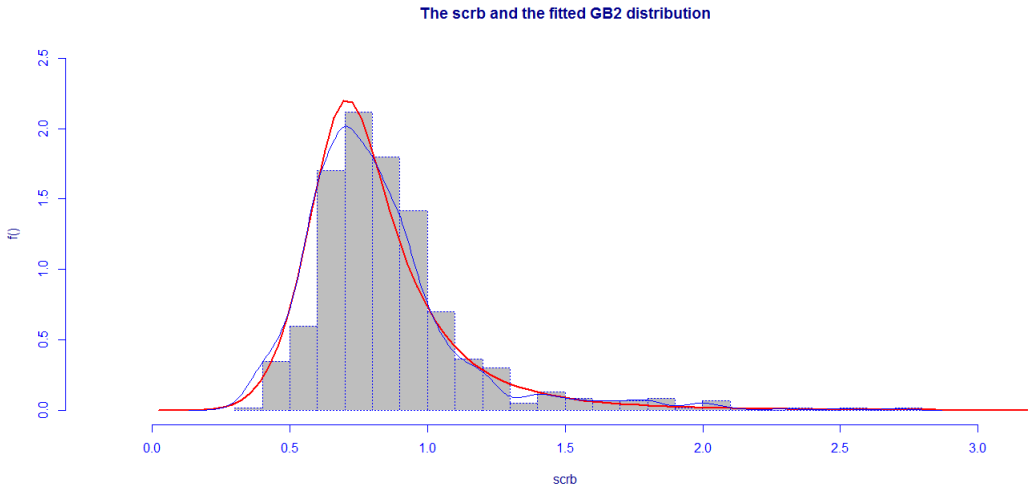

Figura 6.2: Histograma da variável creatinina sérica basal obtido pela função **histDist**(). Linha a vermelha: função densidade paramétrica *GB2*; linha a azul: densidade estimada não-parametricamente.

6.2 Análise do modelo

A construção do modelo é efetuada com base nas variáveis de interesse, idade, sexo, raça ou creatinina no tempo 0, para cada parâmetro da distribuição. No entanto, a variável creatinina sérica no tempo 0, como verificado na análise com o MAG, não tem uma relação linear com a creatinina sérica basal, necessitando da utilização de um suavizador. Uma vez que o *package* GAMLSS disponibiliza diferentes suavizadores, foi necessário escolher um para a variável creatinina sérica no tempo 0. Construíram-se diversos modelos para estimar o parâmetro μ, com apenas uma variável independente, a creatinina sérica no tempo 0 com diferentes suavizadores (disponíveis pelo *package* GAMLSS). Uma vez que os modelos GAMLSS podem ser comparados pelos valores do critério GAIC, escolheu-se o suavizador do modelo univariado que obteve menor valor (Tabela 6.1). O suavizador escolhido foi o *cubic spline* - **scs**(), que será aplicado à

variável creatinina sérica no tempo 0, sempre que esta for selecionada para a estimação de um determinado parâmetro da distribuição.

Tabela 6.1: Modelos univariados com vista à escolha do suavizador para a variável creatinina sérica no tempo 0

Modelo	**Suavizador**	**GAIC**
scrb ∼ cs(scr_t0)	*Cubic spline*	-470.2354
scrb ∼ pvc(scr_t0)	*based on P-splines*	-491.9919
scrb ∼ cy(scr_t0)	*based on P-splines*	-479.0178 - com **warnings**()
scrb ∼ pb(scr_t0)	*based on P-splines*	-491.9913
scrb ∼ ps(scr_t0)	*based on P-splines*	-471.638
scrb ∼ ri(scr_t0)	*based on P-splines*	-375.0739
scrb ∼ fp(scr_t0)	*Fractional Polynomials*	-468.061
scrb ∼ scs(scr_t0)	*Cubic spline*	-497.2156
scrb ∼ lo(scr_t0)	*Loess*	**warnings**()

Para a seleção das variáveis independentes a serem utilizadas na estimação dos parâmetros μ, σ, ν e τ, foram utilizados três métodos de seleção, **stepGAIC**(), **stepGAICAll.A**() e **stepGAICAll.B**(). A **stepGAIC**() é a única função que constrói o modelo GAMLSS individualmente para cada parâmetro da distribuição. O método **stepGAICAll.A**() obteve avisos **warnings**(), revelando problemas na execução do *script*. Assim os seus resultados não serão apresentados neste estudo.

Como o método **stepGAIC**() realiza o ajustamento para cada parâmetro da distribuição individualmente, iniciou-se pelo parâmetro μ, uma vez que Stasinopoulos e Rigby (2015) sugerem que se deve respeitar a hierarquia dos parâmetros. Primeiro deve-se ajustar o modelo para o parâmetro μ, seguido dos parâmetros σ, ν e, por último, o parâmetro τ. A expressão do modelo GAMLSS multivariável obtido pela função **stepGAIC**(), para todos os parâmetros da distribuição, foi a seguinte:

$$g_1(\mu) = \beta_{11}\text{raca} + \beta_{12}\text{sexo} + \beta_{13}\text{idade} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.1}$$

$$g_2(\sigma) = \beta_{22}\text{sexo} + \beta_{23}\text{idade} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.2}$$

$$g_3(\nu) = \text{scs}_{31}(\text{scr\_t0}), \tag{6.3}$$

$$g_4(\tau) = \beta_{41}\text{raca} + \beta_{42}\text{sexo} + \beta_{43}\text{idade}. \tag{6.4}$$

A expressão do modelo GAMLSS anterior, contendo os valores dos coeficientes estimados pelo algoritmo, pode ser escrito na seguinte forma:

$$g_1(\mu) = -0.0813\text{raca} + 0.0483\text{sexo} - 0.0054\text{idade} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.5}$$

$$g_2(\sigma) = -1.07942\text{sexo} + 0.03550\text{idade} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.6}$$

$$g_3(\nu) = \text{scs}_{31}(\text{scr\_t0}), \tag{6.7}$$

$$g_4(\tau) = -0.44286\text{raca} + 0.57384\text{sexo} - 0.03171\text{idade}. \tag{6.8}$$

Os resultados do modelo GAMLSS obtidos encontram-se apresentados na tabela 6.2. Nesta tabela estão representados os valores-p relativos às estimativas das variáveis incluídas no modelo para cada parâmetro da distribuição. Estes valores-p correspondem ao teste de significância do coeficiente, onde é testado se a covariável tem, ou não, influência significativa na variável resposta. Como se pode verificar, alguns dos valores-p são superiores a 0.05, indicando que não é rejeitada a possibilidade de o coeficiente ser igual a zero. Isto indica que a variável independente poderá não ser importante para o modelo de predição. Pela tabela 6.2 é possível concluir que, no modelo obtido para o parâmetro μ, as variáveis raça, idade e sexo poderão ser consideradas pouco relevantes. O mesmo se pode afirmar para a variável raça no modelo obtido para o parâmetro τ.

Tabela 6.2: Modelo GAMLSS - Método **stepGAIC**()

Variável	$\hat{\boldsymbol{\beta}}$	**Valor-p**
Parâmetro - μ		
Raça	-0.0813	0.135
Idade	-0.0054	0.135
Sexo	0.0483	0.410
SCr_T0	0.2380	$< 2 \times 10^{-16}$
Parâmetro - σ		
Idade	0.03550	0.00245
Sexo	-1.07942	0.01364
SCr_T0	-2.12259	1.34×10^{-6}
Parâmetro - ν		
SCr_T0	0.50969	3.37×10^{-8}
Parâmetro - τ		
Raça	-0.44286	0.12490
Idade	-0.03171	0.04618
Sexo	0.57384	0.03579
Global Deviance	-606.001	
Graus de liberdade	24.697	
GAIC	-507.2138	

Ainda para este modelo foram testadas as seguintes interações, sexo*idade, sexo*raca e raca*idade. Como é possível observar pelos resultados dos valores-p obtidos e apresentados na

tabela 6.3, todas as interações foram estatisticamente não significativas. Desta forma, manteve-se as expressões do modelo GAMLSS (6.5, 6.6, 6.7 e 6.8) sem interações.

Tabela 6.3: Interações entre as covariáveis raça, sexo e idade, avaliadas no modelo GAMLSS obtido pelo método **stepGAIC**()

Variável	$\hat{\boldsymbol{\beta}}$	**Valor-p**
Parâmetro - μ		
Idade*Sexo	0.005243	0.3750
Idade*Raça	-0.007807	0.194
Raça*Sexo	-0.043057	0.714
Parâmetro - σ		
Idade*Sexo	-0.02511	0.5109
Parâmetro - τ		
Idade*Sexo	0.01474	0.67032
Idade*Raça	-0.04209	0.286
Sexo*Raça	0.02890	0.9649

Quanto ao método **stepGAICAll.B**() para a seleção das variáveis independentes, é possível observar que, para todos os parâmetros da distribuição, foram inseridas as mesmas covariáveis, sexo e creatinina no tempo 0. A expressão do modelo GAMLSS obtido é a seguinte:

$$g_1(\mu) = \beta_{11}\text{sexo} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.9}$$

$$g_2(\sigma) = \beta_{21}\text{sexo} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.10}$$

$$g_3(\nu) = \beta_{31}\text{sexo} + \text{scs}_{31}(\text{scr\_t0}), \tag{6.11}$$

$$g_4(\tau) = \beta_{41}\text{sexo} + \text{scs}_{41}(\text{scr\_t0}). \tag{6.12}$$

A expressão do modelo GAMLSS anterior, contendo os valores dos coeficientes estimados pelo algoritmo, pode ser escrito na seguinte forma:

$$g_1(\mu) = -0.01710\text{sexo} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.13}$$

$$g_2(\sigma) = 0.11060\text{sexo} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.14}$$

$$g_3(\nu) = -0.1714\text{sexo} + \text{scs}_{31}(\text{scr\_t0}), \tag{6.15}$$

$$g_4(\tau) = 0.11019\text{sexo} + \text{scs}_{41}(\text{scr\_t0}). \tag{6.16}$$

A seleção das mesmas covariáveis para todos os parâmetros da distribuição deve-se ao facto do método **stepGAICAll.B**() obrigar a que todos os parâmetros da distribuição incluam as mesmas covariáveis. Na tabela 6.4 encontram-se os resultados dos valores-p das estimativas dos coeficientes obtidos para cada covariável de cada parâmetro de distribuição.

Tabela 6.4: Modelo GAMLSS - Método **stepGAICAll.B**()

Variável	$\hat{\boldsymbol{\beta}}$	**Valor-p**
Parâmetro - μ		
Sexo	-0.01710	0.642
SCr_T0	0.65611	$< 2 \times 10^{-16}$
Parâmetro - σ		
Sexo	0.11060	0.271
SCr_T0	-3.54282	$< 2 \times 10^{-16}$
Parâmetro - ν		
Sexo	-0.1714	0.185
SCr_T0	1.0108	6.85×10^{-10}
Parâmetro - τ		
Sexo	0.11019	0.537
SCr_T0	1.22473	$< 2 \times 10^{-16}$
Global Deviance	-608.094	
Graus de liberdade	27.00621	
GAIC	-512.0692	

Pela análise dos valores-p, constatou-se que permaneceram no modelo final covariáveis consideradas não significativas. A variável sexo foi considerada não relevante para a estimação de todos os parâmetros da distribuição, enquanto que variável creatinina no tempo 0 é sempre significativa. Uma vez que o modelo obtido pelo método **stepGAIC**() e o obtido pelo método **stepGAICAll.B**() são não aninhados, a medida de avaliação utilizada para comparar os respetivos desempenhos foi o critério GAIC. Este modelo obteve um valor de GAIC de -512.0692 e o modelo obtido pelo método **stepGAIC**() (-507.2138), indicando que o modelo **stepGAICAll.B**() se ajustou melhor aos dados.

6.3 Análise das estimativas

A primeira análise realizada às estimativas obtidas pelos modelos GAMLSS constou da comparação entre os histogramas obtidos a partir dessas estimativas e os obtidos a partir dos valores observados da amostra. No histograma da figura 6.3 - esquerda, encontra-se a representação dos valores da creatinina sérica basal estimados pelo modelo GAMLSS pelo método **stepGAIC**(), enquanto no histograma (Figura 6.3 - direita) os valores observados da creatinina sérica basal. Como é possível observar os histogramas são parecidos, mas com diferenças relevantes, em algumas regiões do gráfico. Por exemplo, o número de indivíduos

ajustados com valores de 1 mg/dl da creatinina basal, são cerca de 300, em comparação com os valores observados, cerca de 250 indivíduos. Para valores superiores a 1.5 mg/dl, o modelo GAMLSS não teve o melhor ajustamento, uma vez que são poucos indivíduos ajustados para esses valores. O mesmo ocorre para valores inferiores a 0.5 mg/dl da creatinina sérica basal, onde o modelo não ajusta nenhum indivíduo com esses valores.

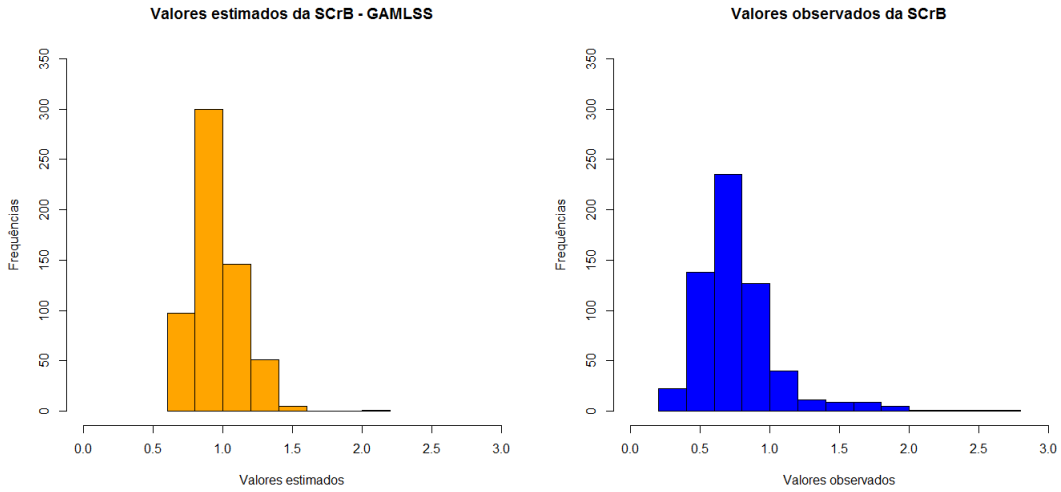

Figura 6.3: Histogramas correspondentes às estimativas obtidas pelo modelo GAMLSS - método **stepGAIC**() (esquerda) e aos valores observados (direita) da creatinina sérica basal.

Os mesmos histogramas foram construídos para as estimativas obtidas pelo modelo GAMLSS - método **stepGAICAll.B**() (Figura 6.4). Como se pode observar, os histogramas são muito diferentes, indicando um mau ajustamento do modelo. Existem alguns valores estimados (Figura 6.4 - esquerda) pelo modelo GAMLSS, extremamente elevados.

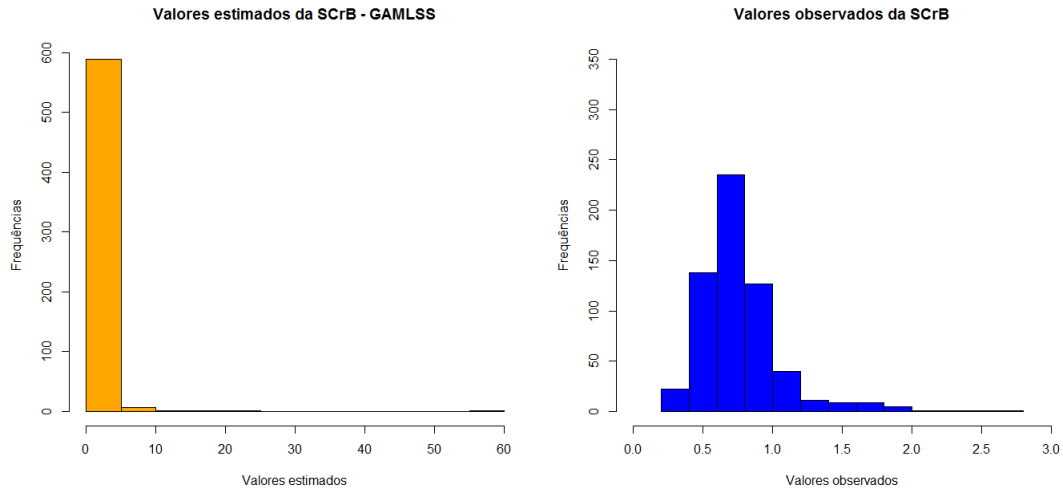

Figura 6.4: Histogramas correspondentes às estimativas obtidas pelo modelo GAMLSS - método **stepGAICAll.B**() (esquerda) e aos valores observados (direita) da creatinina sérica basal.

O gráfico seguinte (Figura 6.5), apresenta, no eixo das abcissas, os valores observados da creatinina sérica basal, e no eixo das ordenadas os valores estimados pelo modelo GAMLSS **stepGAIC**(). A vermelho encontra-se a reta $y = x$. Se o modelo GAMLSS fosse adequado aos dados, a nuvem de pontos (a preto) estaria sobreposta à reta vermelha. No entanto, isso não

acontece e, como é possível observar, o modelo GAMLSS estima diferentes valores de creatinina sérica basal para valores iguais observados. Para valores observados superiores a 1.2 mg/dl, o gráfico revela uma maior dificuldade no ajustamento, uma vez que se observa um desvio bastante acentuado dos pontos em relação à reta $y = x$.

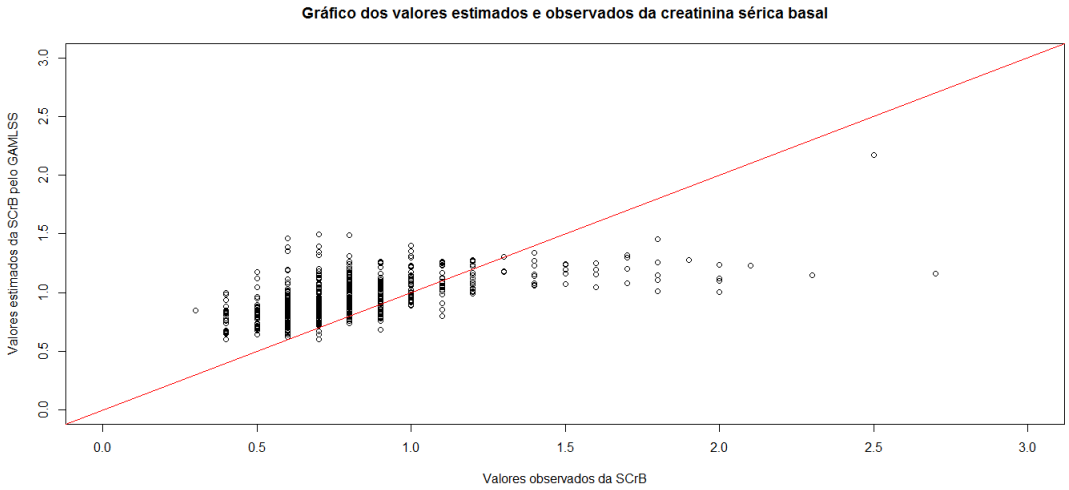

Figura 6.5: Gráfico dos valores observados *versus* os valores estimados pelo modelo GAMLSS - método **stepGAIC**() da creatinina sérica basal. A vermelho encontra-se a reta $y = x$.

Para o modelo GAMLSS estimado, utilizando o método **stepGAICAll.B**(), também foi construído o mesmo gráfico (Figura 6.6). Nesta figura encontram-se dois gráficos, sendo que o inferior representa uma ampliação do gráfico superior, de forma a facilitar a sua visualização. Como era previsto já pelo histograma da figura 6.4, o modelo GAMLSS - método **stepGAICAll.B**() não obteve um bom ajustamento. Alguns valores estimados pelo modelo GAMLSS da creatinina sérica basal foram muito elevados. De facto, foi obtido um valor estimado com cerca de 60 mg/dl e muitos valores entre 5 mg/dl e 25 mg/dl, tendo em conta que o valor máximo observado da creatinina sérica basal é 2.7 mg/dl.

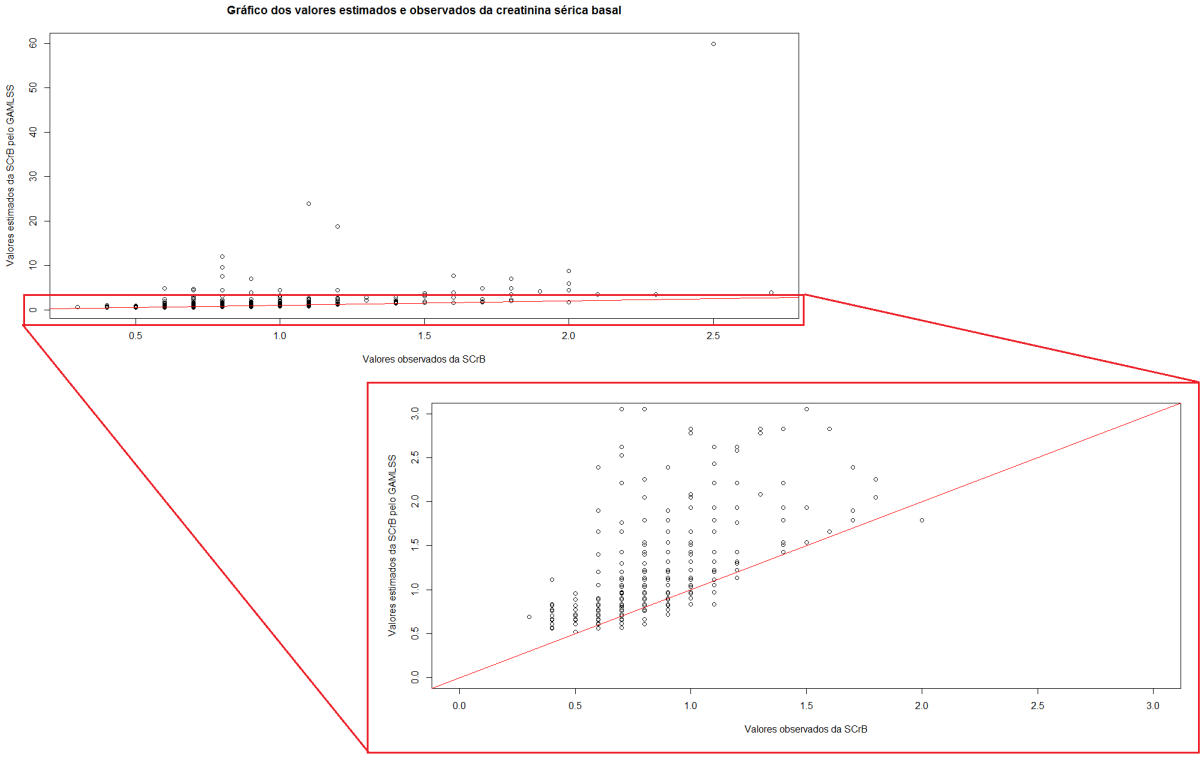

Figura 6.6: Gráfico dos valores observados *versus* os valores estimados, pelo modelo GAMLSS - método **stepGAICAll.B**() da creatinina sérica basal. A vermelho encontra-se a reta $y = x$.

6.4 Análise dos resíduos

Para que o modelo GAMLSS, se ajuste de uma forma adequada, os resíduos *normalised quantile residuals* obtidos terão que ter uma distribuição normal. Assim, foram construídos diagramas em caixa de bigodes e *QQ-plots* para a análise da distribuição dos resíduos.

Na figura 6.7 estão representados os gráficos para análise dos resíduos obtidos pelo modelo GAMLSS - método **stepGAIC**(). Como é possível observar, ambos os gráficos sugerem que a distribuição dos resíduos passa a ser normal. No diagrama em caixa de bigodes (Figura 6.7 - esquerda) são observados algumas observações extremas em ambos os extremos do gráfico. No gráfico *QQ-plot* (Figura 6.7 - direita) também é possível observar alguns resíduos (poucos) com desvios da reta $y = x$, nos extremos do gráfico.

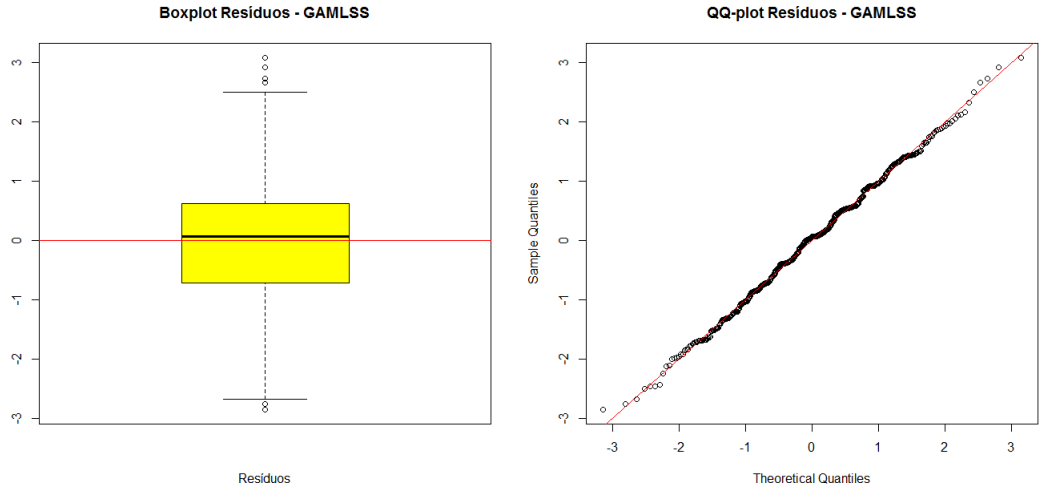

Figura 6.7: Diagrama em caixa de bigodes (esquerda) e *QQ-plot* (direita) dos resíduos obtidos pelo modelo GAMLSS - método **stepGAIC**().

Em relação aos resíduos obtidos pelo modelo GAMLSS do método **stepGAICAll.B**() é possível observar que ambos os gráficos da figura 6.8 sugerem uma distribuição próxima da normal. São observados menos valores extremos no diagrama em caixa de bigodes da figura 6.8 - esquerda, em comparação com o gráfico da figura 6.7 - esquerda. No entanto, pela análise do *QQ-plot*, figura 6.8 - direita parecem existir mais resíduos com maior desvio da reta $y = x$, em comparação com o *QQ-plot* da figura 6.7 - direita.

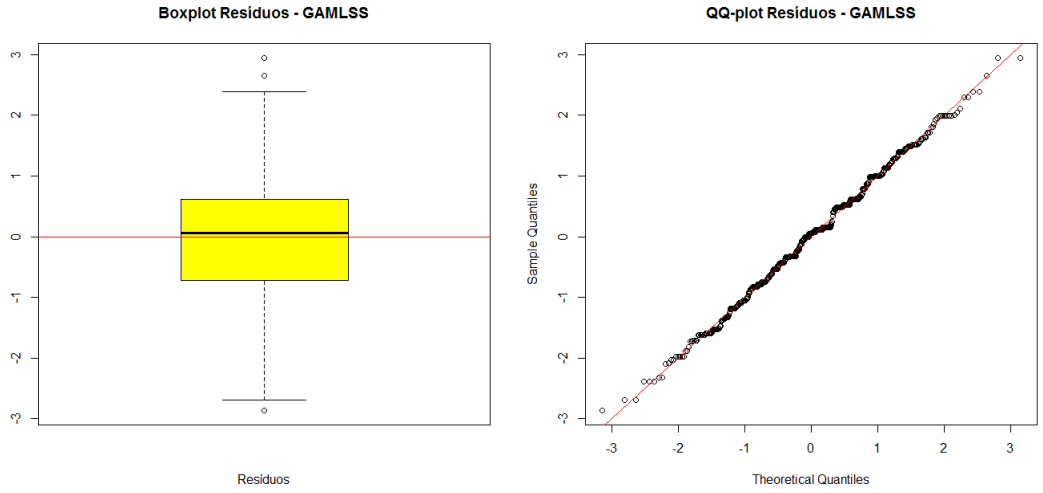

Figura 6.8: Diagrama em caixa de bigodes (esquerda) e *QQ-plot* (direita) dos resíduos obtidos pelo modelo GAMLSS - método **stepGAICAll.B**().

Através da função **plot**() foi possível obter as medidas descritivas dos resíduos obtidos pelos modelos GAMLSS pelo método **stepGAIC**() e pelo método **stepGAICAll.B**(). Estes resultados podem ser visualizados pelos *scripts* da figura 6.9, método **stepGAIC**() e da figura 6.10, método **stepGAICAll.B**(). Os resíduos obtidos pelo método **stepGAIC**() têm um valor média e variância de 0.0022 e de 0.9952, respetivamente, enquanto que os obtidos pelo método **stepGAICAll.B**() têm uma média dos resíduos de -0.00031 e uma variância de 0.999. Os valores para os diferentes métodos foram muito semelhantes, sugerindo que ambos os resíduos

obtidos possam seguir a distribuição da normal (0,1).

```
*******************************************************************
              Summary of the Quantile Residuals
                           mean   =  0.002243666
                       variance   =  0.9951893
                   coef. of skewness  =  -0.02866743
                   coef. of kurtosis  =  2.884111
Filliben correlation coefficient  =  0.998883
*******************************************************************
```

Figura 6.9: Medidas descritivas dos resíduos obtidos pelo modelo GAMLSS - método **stepGAIC**().

```
*******************************************************************
              Summary of the Quantile Residuals
                           mean   =  -0.0003172006
                       variance   =  0.9998826
                   coef. of skewness  =  -0.01073071
                   coef. of kurtosis  =  2.823412
Filliben correlation coefficient  =  0.9985742
*******************************************************************
```

Figura 6.10: Medidas descritivas dos resíduos obtidos pelo modelo GAMLSS - método **stepGAICAll.B**().

Os gráficos obtidos pela função **plot**() constam de quatro gráficos diferentes, '*Against Fitted Values*', '*Against Index*', '*Density Estimate*' e '*Normal Q-Q Plot*'. O último gráfico, *Normal Q-Q Plot* já foi referido para ambos os métodos **stepGAIC**() e **stepGAICAll.B**() representados nas figuras 6.7 - direita e 6.8 - direita, respetivamente.

Na figura 6.11, são apresentados os gráficos obtidos pela função **plot**() do modelo GAMLSS pelo método **stepGAIC**(). O gráfico do canto superior esquerdo representa os resíduos obtidos *versus* os valores estimados da creatinina sérica basal pelo modelo GAMLSS. Já o gráfico do canto superior direito, da mesma figura, representa o *index* de cada indivíduo da amostra *versus* os resíduos obtidos pelo mesmo modelo. Para ambos os gráficos não é observado qualquer tipo de padrão dos resíduos, indicando um bom ajustamento do modelo. O terceiro gráfico, localizado no canto inferior esquerdo da figura 6.11, é um gráfico de estimação não-paramétrica (suavizador de Kernel) da densidade dos resíduos do modelo. Este gráfico também indica um bom ajustamento do modelo, uma vez que tem uma forma semelhante à da função densidade da normal (0,1).

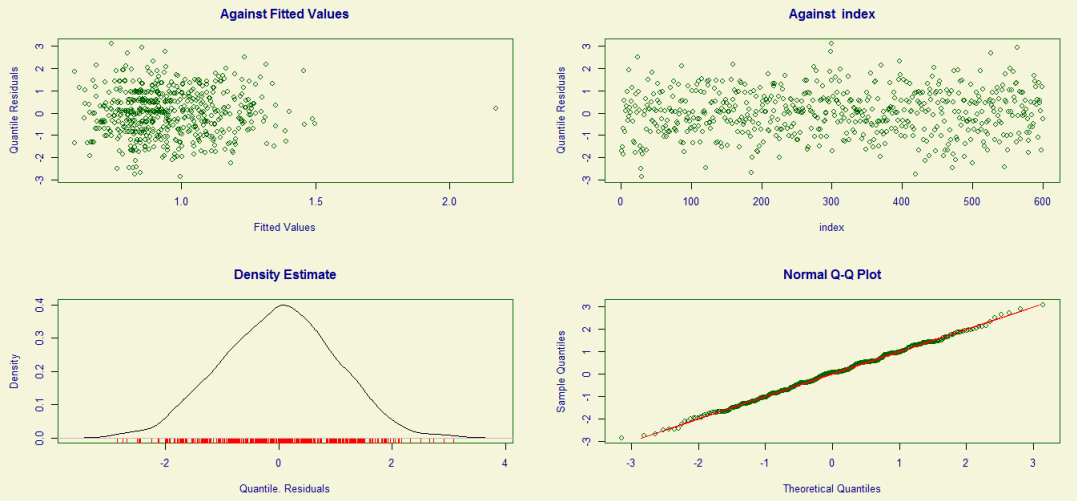

Figura 6.11: Gráfico dos resíduos obtidos pelo modelo GAMLSS - método **stepGAIC**(), através da função **plot**().

A mesma função **plot**() também foi aplicada ao modelo GAMLSS pelo método **stepGAICAll.B**() (Figura 6.12). Como os valores estimados da creatinina sérica basal foram maus para alguns indivíduos, é de esperar que o gráfico '*Against Fitted Values*' não seja o ideal. Esta conclusão pode ser observada no gráfico do canto superior esquerdo da figura 6.12, onde é possível observar uma concentração de resíduos à esquerda. Em relação ao restantes gráficos da figura 6.12, pode-se concluir que os resíduos têm um comportamento semelhante aos obtidos pelo modelo GAMLSS com o método **stepGAIC**().

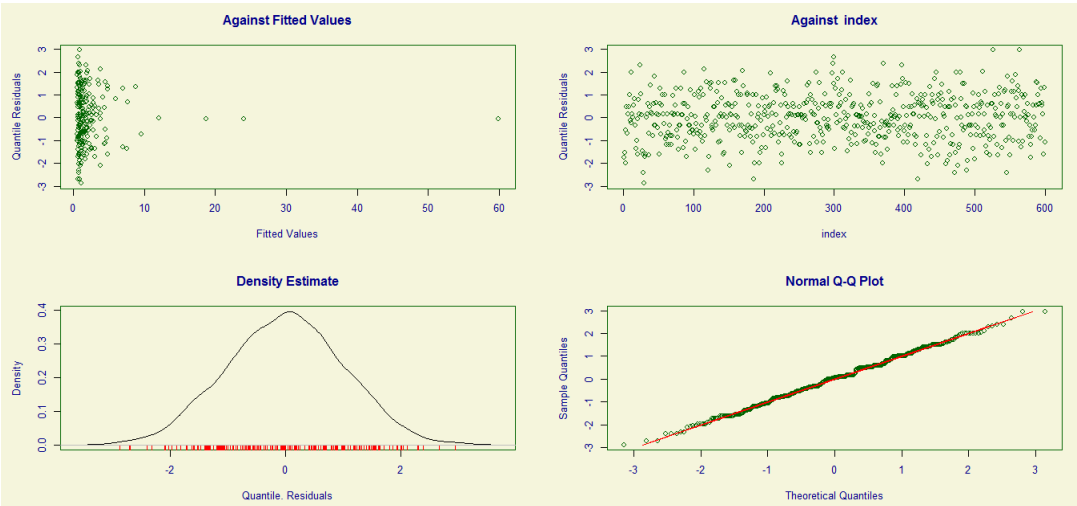

Figura 6.12: Gráfico dos resíduos obtidos pelo modelo GAMLSS - método **stepGAICAll.B**(), através da função **plot**().

Outra forma de analisar os resíduos é através do *Worm plot*. Pelo gráfico do modelo GAMLSS - método **stepGAIC**() (Figura 6.13) é possível observar que os pontos não se encontram muito longe da reta $y = 0$. Isto indica um bom ajustamento do modelo. Existem apenas alguns resíduos nos extremos que apresentam uma maior discrepância no que diz respeito à distância à reta $y = 0$. Além disso, apenas um dos resíduos encontra-se dentro do semi-círculo superior, sendo uma boa indicação acerca do ajustamento. A curva a vermelho representa o ajustamento

cúbico aos resíduos. O facto de não refletir nenhum padrão semelhante aos da figura 3.2 é indicador de um modelo bem ajustado aos dados.

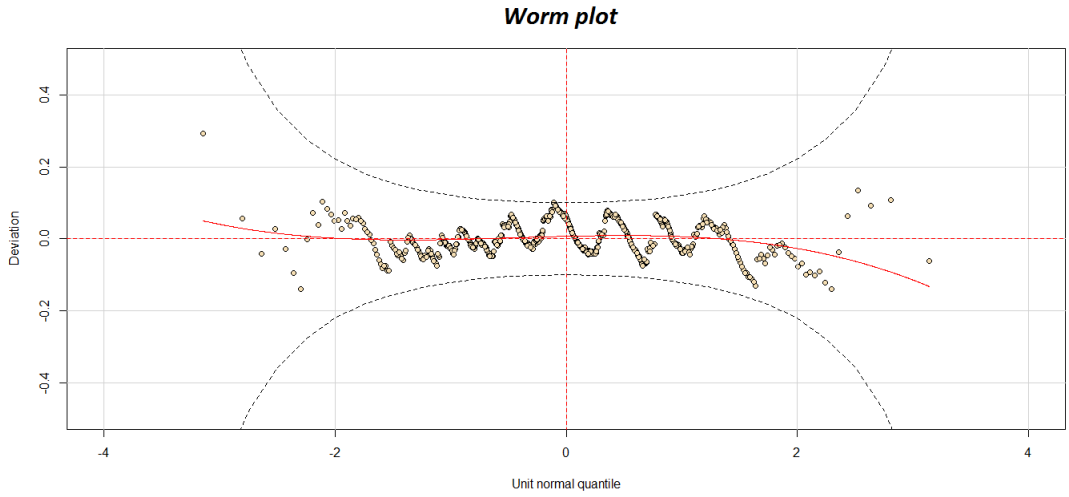

Figura 6.13: *Worm plot* dos resíduos obtidos pelo modelo GAMLSS - método **stepGAIC**().

O *Worm plot* dos resíduos obtidos pelo modelo GAMLSS utilizando o método **stepGAICAll.B**() (Figura 6.14) obteve piores resultados, quando comparados com os da figura 6.13. Através do gráfico é possível observar que os resíduos apresentam um maior afastamento da reta $y = 0$. Além disso, é possível observar alguns resíduos sobrepostos sobre os semi-círculos inferior e superior.

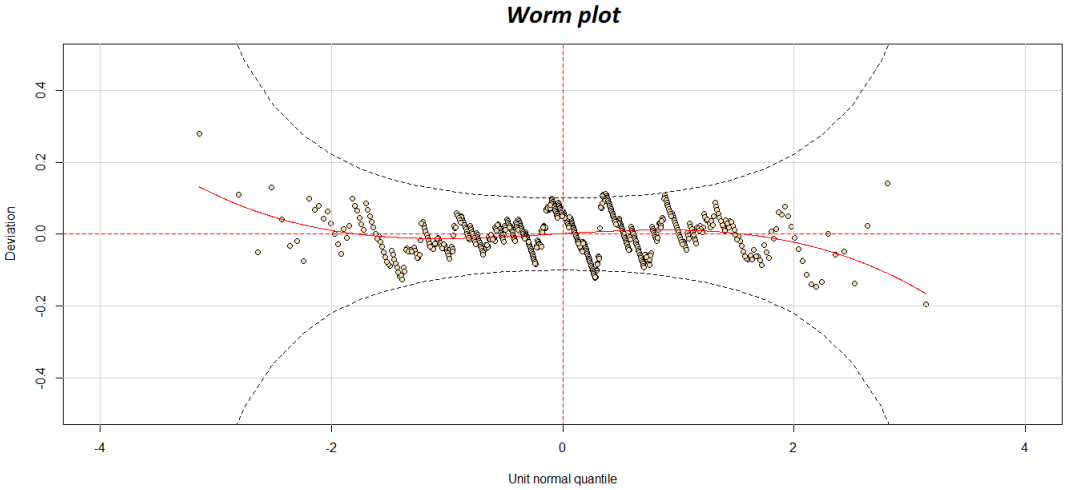

Figura 6.14: *Worm plot* dos resíduos obtidos pelo modelo GAMLSS - método **stepGAICAll.B**().

Devido aos problemas de ajustamento do modelo GAMLSS pelo método **stepGAICAll.B**(), optou-se por escolher o modelo GAMLSS obtido pelo **stepGAIC**() para realizar uma análise mais extensa aos dados. Assim, embora os resíduos obtidos por este modelo indiquem um bom ajustamento do modelo, os valores estimados não foram os esperados. Desta forma, voltou a realizar-se a análise dos dados sem as seis observações mais extremas identificadas no diagrama em caixa de bigodes da figura 6.7 - esquerda.

6.5 Análise dos dados sem os valores extremos

Os valores dos resíduos mais extremos obtidos pelo modelo GAMLSS - método **stepGAIC**(), identificados no diagrama em caixa de bigodes da figura 6.7 (esquerda), foram eliminados da base de dados original, com o intuito de analisar o comportamento do modelo nestas circunstâncias. Foi seguidamente realizada uma nova análise, utilizando os modelos GAMLSS. Primeiramente aplicou-se a função **fitDist**() para obter a função de distribuição mais adequada à creatinina sérica basal. Obteve-se a mesma distribuição do primeiro modelo GAMLSS, *GB2 - generalized beta type 2*, como se pode constatar pelo *script* da figura 6.15. Construiu-se um novo histograma (Figura 6.16) utilizando a função **histDist**(), onde as constantes dos parâmetros da distribuição obtidos foram $\hat{\mu} = -0.3534$, $\hat{\sigma} = 9.679$, $\hat{\nu} = -0.3997$ e $\hat{\tau} = -0.8218$. Uma vez que a distribuição da variável dependente é a mesma, *GB2*, as funções de ligação são as mesmas: μ, ν e τ, função logarítmica e σ, função identidade.

```
> fitting_dist_out<-fitDist(dados_finais_sem_out_gamlss[,1],
+ type=c('realplus'),extra=c('BCPE','BCCG','BCT','GB2', 'WEI','WEI2'))
Warning message:
In MLE(ll2, start = list(eta.mu = eta.mu, eta.sigma = eta.sigma),  :
  possible convergence problem: optim gave code=1 false convergence (8)
> fitting_dist_out$fits
      GB2      BCTo       BCT     BCPEo      BCPE    exGAUS      BCCG     BCCGo
-36.20149 -32.65588 -32.65588 -32.60748 -32.60748 -23.39050 -15.23989 -15.23989
       GG    IGAMMA       GIG     LOGNO        IG        GA      WEI2      WEI3
-13.31773 -13.31222 -11.31222  10.81049  15.84424  65.25424 245.46462 245.46462
      WEI       EXP   PARETO2
245.46462 954.28906 956.28929
> fitting_dist_out$failed
list()
```

Figura 6.15: *Script* e respetivo *output* da função **fitDist**() aplicada à variável creatinina sérica basal sem os valores extremos.

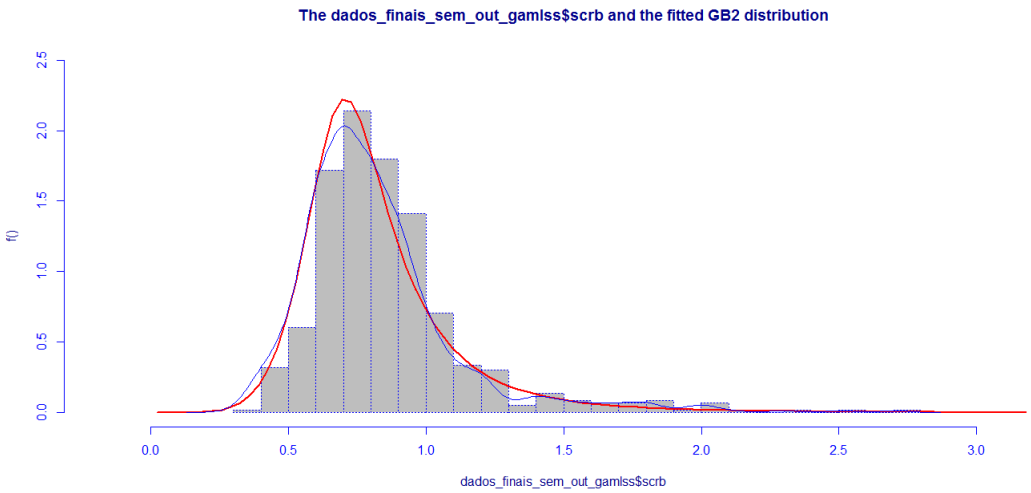

Figura 6.16: Histograma da variável creatinina sérica basal obtido pela função **histDist**(), sem os valores extremos. Linha a vermelha: função densidade paramétrica *GB2*; linha a azul: densidade estimada não-parametricamente.

Após a seleção da distribuição para a variável dependente, foi necessário selecionar as variáveis independentes para cada parâmetro da distribuição. Apenas foi utilizado o método **stepGAIC**(), uma vez que foi o que obteve melhores resultados, como já referido anteriormente.

As variáveis selecionadas para cada parâmetro da distribuição foram iguais às do primeiro modelo GAMLSS implementado com base no método **stepGAIC**(), exceto para o parâmetro μ. Para este parâmetro da distribuição, a variável sexo não foi inserida. Assim, as expressões do novo modelo GAMLSS são as seguintes:

$$g_1(\mu) = \beta_{11}\text{raca} + \beta_{13}\text{idade} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.17}$$

$$g_2(\sigma) = \beta_{22}\text{sexo} + \beta_{23}\text{idade} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.18}$$

$$g_3(\nu) = \text{scs}_{31}(\text{scr\_t0}), \tag{6.19}$$

$$g_4(\tau) = \beta_{41}\text{raca} + \beta_{42}\text{sexo} + \beta_{43}\text{idade}. \tag{6.20}$$

A expressão do modelo GAMLSS anterior, contendo os valores dos coeficientes estimados pelo algoritmo, pode ser escrito na seguinte forma:

$$g_1(\mu) = -0.08125\text{raca} - 0.0043\text{idade} + \text{scs}_{11}(\text{scr\_t0}), \tag{6.21}$$

$$g_2(\sigma) = -0.7638\text{sexo} + 0.0283\text{idade} + \text{scs}_{21}(\text{scr\_t0}), \tag{6.22}$$

$$g_3(\nu) = \text{scs}_{31}(\text{scr\_t0}), \tag{6.23}$$

$$g_4(\tau) = -0.5281\text{raca} + 0.3191\text{sexo} - 0.0279\text{idade}. \tag{6.24}$$

As estimativas dos coeficientes, valores-p e outras medidas resultantes do ajustamento do modelo GAMLSS, encontram-se na tabela 6.5. Como é possível observar, este modelo obteve alguns valores-p não significativos para os testes de significância dos coeficientes β. Como no primeiro modelo GAMLSS - método **stepGAIC**(), para as variáveis raça e idade obtém-se um valor-p superior a 0.05, para o parâmetro μ, indicando que não devem ser inseridas no modelo. Todos os coeficientes das variáveis incluídas para o parâmetro σ e ν tiveram valores-p significativos. No entanto, o parâmetro τ apenas obteve o coeficiente da variável sexo significativo. Em comparação com o primeiro modelo GAMLSS - método **stepGAIC**(), apenas a variável raça para o parâmetro τ não tinha um valor-p inferior a 0.05, mas os valores-p para as variáveis raça e idade são próximos de 0.05. Sendo assim, embora este parâmetro não rejeite a não-significância destes dois coeficientes, os seus valores-p estão muito próximos da rejeição, ou seja, próximos de 0.05.

O valor de GAIC obtido para este modelo (Tabela 6.5) foi menor do que o modelo GAMLSS obtido pelo método **stepGAIC**(), sugerindo que este poderá ter melhor desempenho.

Tabela 6.5: Modelo GAMLSS - dados sem os valores extremos

Variável	$\hat{\boldsymbol{\beta}}$	**Valor-p**
Parâmetro - μ		
Raça	-0.08125	0.118
Idade	-0.0043	0.209
SCr_T0	0.2253	$< 2 \times 10^{-16}$
Parâmetro - σ		
Idade	0.0283	0.00848
Sexo	-0.7638	0.02174
SCr_T0	-2.3225	6.42×10^{-7}
Parâmetro - ν		
SCr_T0	0.5511	1.98×10^{-7}
Parâmetro - τ		
Raça	-0.5281	0.066776
Idade	-0.0279	0.066960
Sexo	0.3191	0.000534
Global Deviance	-646.5825	
Graus de liberdade	24.33838	
GAIC	-549.229	

Na figura 6.17, encontram-se os histogramas dos valores estimados pelo modelo GAMLSS da creatinina sérica basal (esquerda) e no histograma da direita encontram-se os valores observados da creatinina sérica basal. Comparando os histogramas da figura 6.17, é possível observar algumas semelhanças, assim como foi observado nos histogramas da figura 6.3. No entanto, se compararmos o histograma da figura 6.3 - esquerda, e o histograma da figura 6.17 - esquerda é possível observar uma melhoria na predição dos valores da creatinina sérica basal, uma vez que há uma diminuição do número de indivíduos para valores de creatinina sérica basal igual a 1 mg/dl.

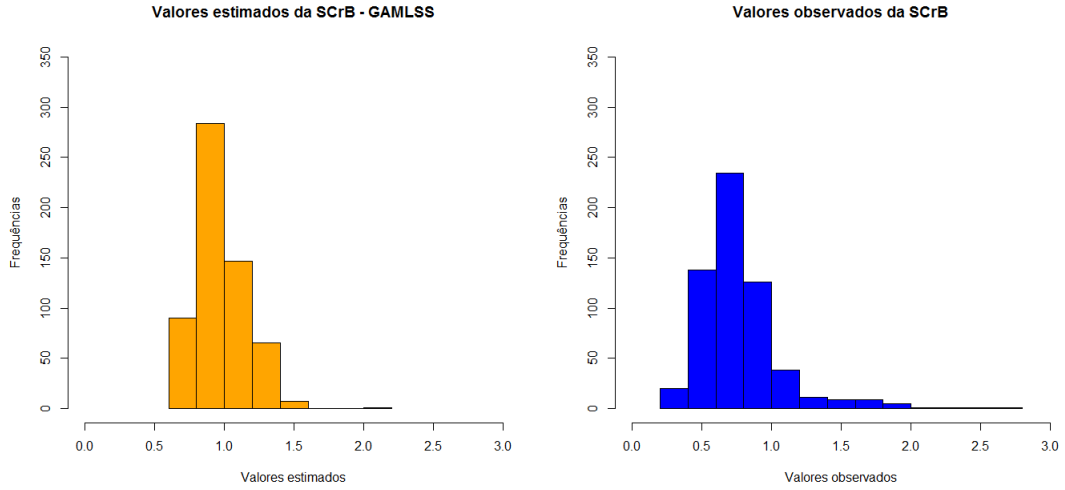

Figura 6.17: Histogramas dos valores observados (direita) e estimados (esquerda) da creatinina sérica basal através do modelo GAMLSS obtido pelo método **stepGAIC**, considerando os resíduos sem os valores extremos.

O gráfico dos valores estimados pelo modelo GAMLSS *versus* os valores observados da creatinina sérica basal, (Figura 6.18) é muito semelhando ao obtido pelo primeiro modelo GAMLSS (Figura 6.5). Os valores obtidos pelo modelo GAMLSS, superiores a 1.2 mg/dl mantêm o desvio em relação à reta a vermelho, $y = x$.

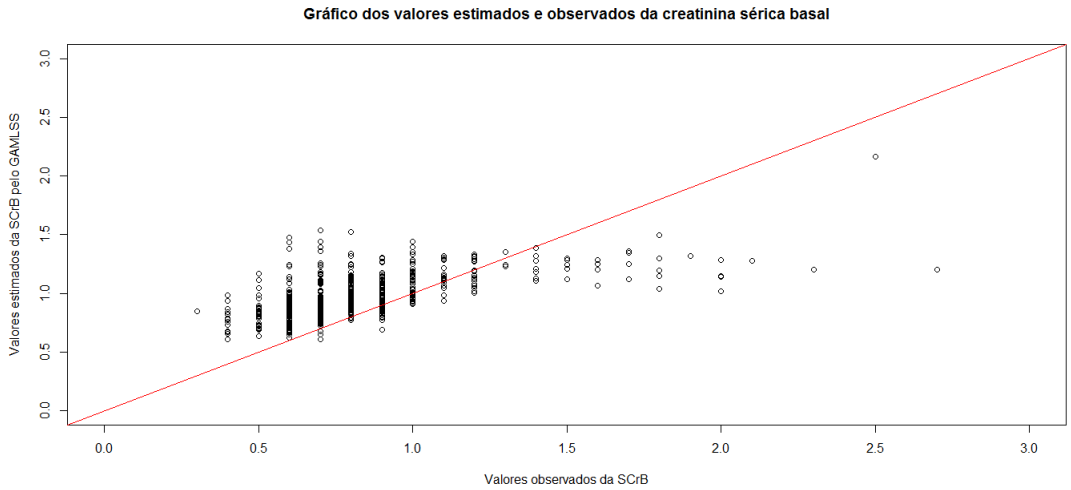

Figura 6.18: Gráfico dos valores observados *versus* os valores estimados pelo modelo GAMLSS, da creatinina sérica basal, considerando os resíduos sem os valores extremos.

Na análise dos resíduos manteve-se a mesma abordagem que a nos modelos anteriores. Na figura 6.19 - esquerda, encontra-se o diagrama em caixa de bigodes e à direita o *QQ-plot* dos resíduos obtidos pelo modelo GAMLSS, dos dados sem os valores extremos. No diagrama em caixa de bigodes, já não é possível observar estes valores extremos, identificados no gráfico 6.7 - esquerda. O gráfico *QQ-plot* 6.19 - direita obtido não mostra grandes variações em relação ao gráfico 6.7 - direita. Desta forma, parece manter-se o mesmo comportamento dos resíduos quando comparados com os obtidos pelo primeiro modelo GAMLSS - método **stepGAIC**().

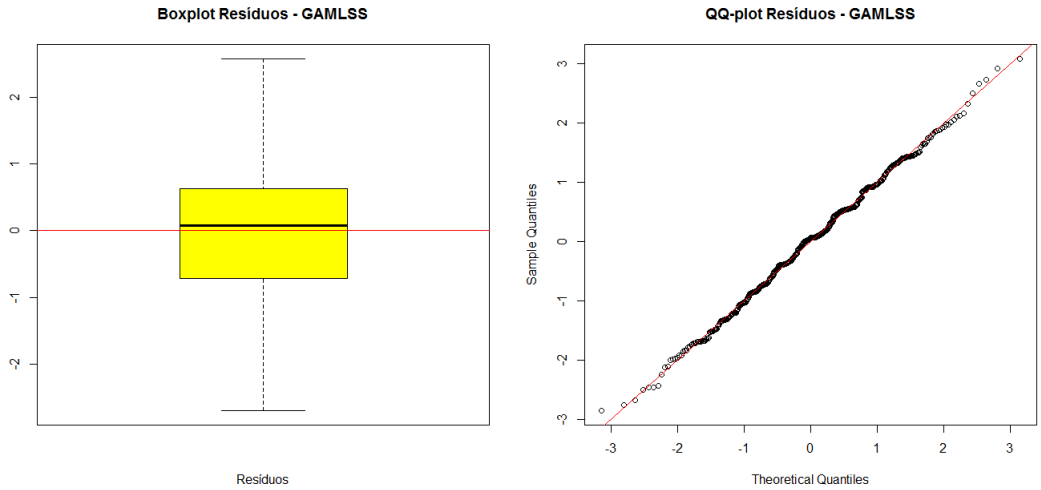

Figura 6.19: Diagrama em caixa de bigodes (esquerda) e *QQ-plot* (direita) dos resíduos obtidos pelo modelo GAMLSS, considerando os resíduos sem os valores extremos.

O *script* obtido pela função **plot**() aplicado aos resultados obtidos pelo modelo GAMLSS é apresentado na figura 6.20. A média dos resíduos obtidos pelo modelo GAMLSS foi de 0.0015, enquanto o valor da variância foi de 0.9960. Estes valores mantêm-se muito semelhantes a qualquer um dos modelos GAMLSS ajustados anteriormente.

```
*******************************************************************
              Summary of the Quantile Residuals
                           mean   =  0.00158933
                       variance   =  0.9960658
                   coef. of skewness  =  -0.04868385
                   coef. of kurtosis  =  2.639592
Filliben correlation coefficient  =  0.9981787
*******************************************************************
```

Figura 6.20: Medidas descritivas dos resíduos obtidos pelo modelo GAMLSS, considerando os resíduos sem os valores extremos.

Os gráficos '*Against Fitted Values*' e '*Against index*' obtidos pela função **plot**(), observados na figura 6.21, não mostram nenhum padrão, indicando um bom ajustamento do modelo. No entanto, é possível observar um resíduo mais distante no gráfico '*Against Fitted Values*'. Este resíduo, já identificado no primeiro modelo GAMLSS pelo método **stepGAIC**(), permaneceu mesmo depois da eliminação dos valores dos resíduos extremos. Em relação ao gráfico '*Density Estimate*' e '*Normal Q-Q Plot*' da figura 6.21, mantém-se a sugestão de que os resíduos obtidos são provenientes de uma população com distribuição normal.

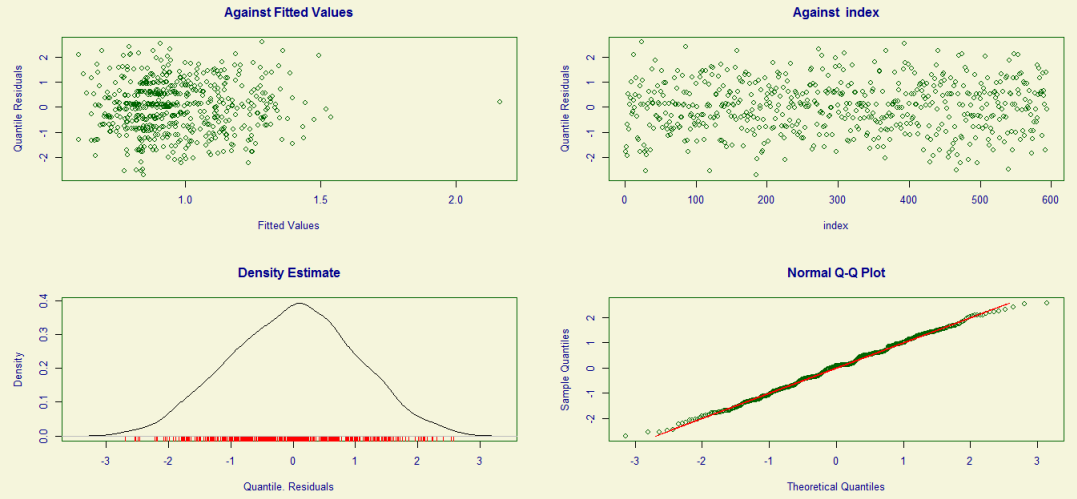

Figura 6.21: Gráficos dos resíduos obtidos pelo modelo GAMLSS sem os valores extremos, através da função **plot**().

Quando os resíduos são analisados através do *Worm plot* é possível observar uma alteração da distribuição em relação ao *Worm plot* da figura 6.13. A curva de ajustamento cúbico afastou-se da reta $y = 0$, ganhando um comportamento mais oblíquo, em forma de 'S'. Embora nenhum dos resíduos se encontre dentro dos semi-círculos, este gráfico poderá indicar uma falha na modelação da curtose.

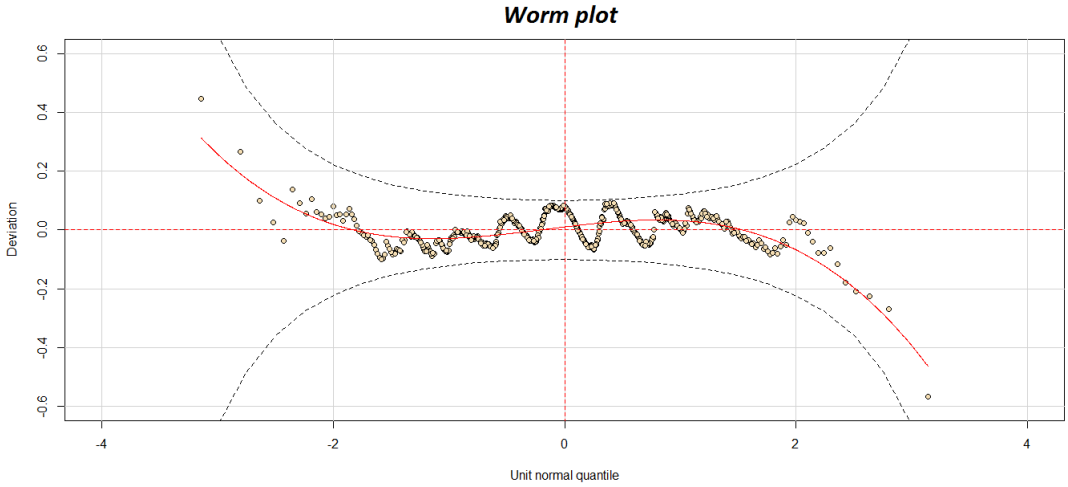

Figura 6.22: *Worm plot* dos resíduos obtidos pelo modelo GAMLSS, sem os valores extremos.

Capítulo 7

Discussão

Identificar uma metodologia que permita estimar a creatinina sérica basal é de elevada importância para a prática clínica, no que diz respeito ao diagnóstico da lesão renal aguda. De facto, um diagnóstico precoce desta doença pode prevenir patologias mais complicadas com necessidade de tratamentos invasivos ou até mesmo levar à morte. Para o diagnóstico da lesão renal aguda é necessário ter o conhecimento do valor da creatinina sérica basal, dado este valor ser utilizado nos critérios de classificação da doença. Uma vez que este valor nem sempre está disponível é necessário recorrer a métodos de cálculo para a sua determinação.

Através de uma revisão da literatura sobre a lesão renal aguda foi possível identificar quais as variáveis de importância clínica a serem utilizadas neste estudo. Desta forma, apenas a creatinina sérica basal, a idade, a raça, o sexo e a creatinina sérica no tempo 0 foram analisadas.

A abordagem escolhida para atingir o objetivo deste estudo incluiu a utilização das técnicas de regressão. Assim, iniciou-se a análise dos dados pelos habituais modelos de regressão lineares generalizados, dada a distribuição contínua da variável resposta. Para ajustar este modelo foram consideradas todas as covariáveis (idade, sexo, raça e creatinina sérica no tempo 0). No entanto, a adequabilidade do ajustamento não foi a melhor dada a não normalidade dos resíduos obtidos.

O método de análise seguinte constou da utilização de modelos aditivos generalizados para estimar a creatinina sérica basal em função das variáveis idade, sexo, raça e creatinina no tempo 0. A vantagem destes modelos, quando comparados com os MLG, tem a ver com o facto de poderem ser utilizadas funções suavizadoras para modelar a relação existente entre as covariáveis contínuas e a variável resposta, conferindo-lhe assim uma maior flexibilidade. No entanto, de novo foi violada a condição de aplicabilidade no que diz respeito à normalidade dos resíduos. Desta forma, foi necessário recorrer a outro método de análise, para tentar identificar um modelo que melhor se ajuste aos dados.

Uma vez que os modelos MLG e MAG não obtiveram os melhores resultados, recorreu-se, numa análise seguinte, à utilização de modelos GAMLSS. Esta nova técnica de regressão foi mais promissora devido à flexibilidade que tem na escolha da distribuição da variável resposta. Através da análise dos dados, utilizando esta metodologia, concluiu-se que a distribuição da variável resposta não é, de facto, uma distribuição normal. A distribuição selecionada para

a creatinina sérica basal foi a distribuição beta generalizada tipo 2. No entanto, quando se utilizou a função disponível no *package* GAMLSS, para a seleção da distribuição, obteve-se mensagens de *Warnings*, que poderão indicar possíveis erros de convergência na sua execução. Mesmo assim, prosseguiu-se com a análise dos dados utilizando essa distribuição, uma vez que nos exemplos práticos apresentados por Stasinopoulos et al. (2015) também apresentavam as mesma mensagens de *Warnings* e os autores prosseguiram com a análise dos modelos GAMLSS.

No ajustamento do modelo foram considerados os quatro parâmetros da distribuição GB2 (valor médio, variância, assimetria e curtose) que, se por um lado podem dificultar a convergência do modelo GAMLSS, por outro, atribuem uma maior flexibilidade à forma da função densidade.

Como a relação entre a creatinina sérica basal e a covariável creatinina sérica no tempo 0 não é linear, manteve-se o suavizador para esta variável independente, já utilizado aquando da modelação dos dados com os MAG. Embora tenha sido selecionada e utilizada na análise dos GAMLSS, a função de suavização **scs**(), através do critério GAIC, Stasinopoulos et al. (2015), recomendam a utilização do suavizador **pb**(), uma vez que este estima o parâmetro de suavização utilizando a máxima verosimilhança local.

De facto, o *software* que permite a implementação dos GAMLSS está muito completo, permitindo determinar o melhor valor do parâmetro de suavização através da função **find.hyper**(). No entanto, nunca foi possível obter resultados, porque a função nunca conseguiu ser executada.

O método **stepGAIC**(), utilizado na seleção das covariáveis para os diferentes parâmetros da distribuição, ajusta um modelo para cada parâmetro individualmente. Quando **stepGAIC**() ajusta um modelo para um determinado parâmetro da distribuição (μ, σ, ν e τ), este determina as medidas de desempenho considerando que os outros parâmetros da distribuição são modelados tendo em conta que todas as covariáveis estão incluídas. Desta forma, poderá ocorrer uma má seleção das covariáveis a considerar na modelação de cada parâmetro. Esta poderá ser a razão pela qual os valores-p, correspondentes aos testes de significância, de alguns coeficientes de regressão sejam superiores a 0.05. Uma possível solução para este problema passa pela utilização do método **stepGAICAll.A**() que seleciona as covariáveis para todos os parâmetros da distribuição em simultâneo. Durante este estudo nunca foi possível obter resultados com este método, uma vez que não se obteve convergência da função **stepGAICAll.A**(). Uma vez que o procedimento de estimação dos quatro parâmetros da distribuição é feito em simultâneo, antecipamos que a amostra de 600 indivíduos possa ser demasiado pequena, provocando a não convergência do algoritmo.

O método **stepGAICAll.B**() é menos flexível que o método **stepGAICAll.A**(), porque seleciona as mesmas covariáveis para todos os parâmetros da distribuição. Se compararmos os valores GAIC dos modelos obtidos pelo método **stepGAICAll.B**() e pelo obtido com o método **stepGAIC**(), o primeiro é o que apresenta melhores resultados, ou seja, com valor menor GAIC. Pela análise do comportamento dos resíduos obtidos por ambos os métodos não foi possível observar grandes diferenças, não sendo possível escolher um melhor modelo

GAMLSS. No entanto, pela análise das estimativas obtidas pelos modelos, as do método **stepGAICAll.B**() demonstraram um mau ajustamento do modelo, embora tendo um menor valor de GAIC. Sendo assim, o modelo GAMLSS pelo método **stepGAIC**() foi o escolhido, uma vez que obteve melhores resultados que o modelo GAMLSS com o método **stepGAICAll.B**().

Na análise seguinte, utilizando o modelo GAMLSS obtido pelo método **stepGAIC**(), identificou-se os resíduos com valores extremos no diagrama em caixa de bigodes (Figura 6.7 - esquerda) e eliminou-se esses indivíduos da amostra. Efetuou-se a mesma análise GAMLSS, onde se obteve a mesma distribuição, a GB2, para a variável creatinina sérica basal. Para a seleção das covariáveis, apenas foi utilizado o método **stepGAIC**(), uma vez que nos modelos anteriores o método **stepGAICAll.A**() nunca conseguiu ser executado e o **stepGAICAll.B**() obteve maus valores estimados da SCr basal. O modelo GAMLSS obtido, utilizando a amostra sem os valores extremos, foi aquele que obteve melhores resultados de entre todos os modelos GAMLSS. De facto, o valor do critério GAIC foi menor em comparação com os restantes modelos, mantendo os resíduos uma distribuição normal.

Através do *Worm plot* obtido pelo último modelo GAMLSS (Figura 6.22), este sugere uma falha na modelação do parâmetro de distribuição curtose. Assim, para uma futura análise utilizando os modelos GAMLSS, será necessário ter em atenção à modelação deste parâmetro.

Uma alternativa para a eventual melhoria da qualidade das estimativas da creatinina sérica basal passa pela estimação de um modelo GAMLSS para cada raça. Uma vez que na literatura os valores da creatinina sérica podem ser influenciados pela raça, é aconselhado realizar a mesma abordagem ajustando diferentes modelos GAMLSS para cada uma das raças. De facto esta é reconhecida como uma variável que, além do género, influencia os valores da creatinina sérica basal, como referido na literatura.

Embora as medidas de diagnóstico indiquem que os modelos GAMLSS estão bem ajustados, os valores estimados não foram os desejados. No entanto, em comparação com os modelo MLG e MAG, os modelos GAMLSS conseguiram ultrapassar a falha da condição de aplicabilidade relativa à normalidade dos resíduos.

Capítulo 8

Conclusão

A não verificação das condições de aplicabilidade dos modelos de regressão põem em causa a qualidade das estimativas obtidas. Assim, para os MLG, onde não foi possível obter a normalidade dos resíduos, conclui-se que o respetivo modelo não poderia ter uma boa capacidade de predição da variável resposta. Na tentativa de alcançar a normalidade dos resíduos, aplicou-se uma transformação logarítmica à variável resposta não tendo, no entanto, havido uma normalização dos resíduos.

Outra abordagem considerada, recorreu à transformação de Box- Cox, mas nunca foi possível obter a estimativa do seu parâmetro de transformação.

A abordagem seguinte utilizou os MAG para modelar os dados, uma vez que a relação entre a variável dependente e as independentes poderia ser não linear. Ainda assim, os resíduos obtidos por este modelo não verificaram as condições de aplicabilidade.

O principal objetivo deste estudo foi a aplicação dos modelos GAMLSS, devido à maior flexibilidade na escolha da distribuição da variável resposta, com o intuito de conduzir à normalidade da distribuição dos resíduos. Através dos resultados obtidos pelo modelo GAMLSS ficou provada a normalidade dos resíduos. No entanto, este modelo GAMLSS não pode ser considerado 'ideal', uma vez que as suas estimativas não se revelaram as melhores, quando comparadas com os valores observados da amostra. Neste contexto, será necessário realizar novas análises utilizando os GAMLSS, uma vez que existiram problemas de execução computacional em algumas das funções utilizadas. Concluiu-se, assim, que o modelo obtido não será o mais adequado para utilizar, na prática clínica, para estimar a creatinina sérica basal, tão necessária no auxílio do diagnóstico da lesão renal aguda.

Referências Bibliográficas

Abensur, H. (2011). *Biomarcadores na Nefrologia*. Sociedade Brasileira de Nefrologia.

Bagshaw, S. M., Uchino, S., Cruz, D., Bellomo, R., Morimatsu, H., Morgera, S., Schetz, M., Tan, I., Bouman, C., Macedo, E., Gibney, N., Tolwani, A., Straaten, H. M. O.-v., Ronco, C., and Kellum, J. A. (2009). A comparison of observed versus estimated baseline creatinine for determination of RIFLE class in patients with acute kidney injury. *Nephrol Dial Transplant*, 24:2739–2744.

Eaton, D. C. and Pooler, J. P. (2009). *Vander's Renal Physiology*. Mc Graw Hill - Medical, 7 edition.

Florencio, L. D. E. A. (2010). *Engenharia de Avaliações com Base em Modelos GAMLSS*. Master thesis, Universidade Federal de Pernambuco.

Gaião, S. and Cruz, D. N. (2010). Baseline creatinine to define acute kidney injury: is there any consensus? *Nephrol Dial Transplant*, 25:3812–3814.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models - A roughness penalty approach*. Chaoman & Hall/CRC.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, 1 edition.

Hsu, J., Johansen, K. L., Hsu, C.-y., Kaysen, G. A., and Chertow, G. M. (2008). Higher serum creatinine concentrations in black patients with chronic kidney disease: beyond nutritional status and body composition. *Clinical Journal of the American Society of Nephrology*, 4(3):992–997.

Kumar, P. and Clark, M. (2009). *Clinical Medicine*. Saunders - Elsevier, 7 edition.

Leung, K. C. W., Tonelli, M., and James, M. T. (2013). Chronic kidney disease following acute kidney injury , risk and outcomes. *Nature Reviews Nephrology*, 9:77–85.

Levey, A. S., Becker, C., and Inker, L. A. (2015). Glomerular Filtration Rate and Albuminuria for Detection and Staging of Acute and Chronic Kidney Disease in Adults: A Systematic Review. *JAMA*, 313(8):837–846.

Longo, D. L., Kasper, D. L., Jameson, J. L., Fauci, A. S., Hauser, S. L., and Loscalzo, J. (2012). *Harrison's Principles of Internal Medicine*. McGraw-Hill - Medicine, 18 edition.

Osborne, J. W. (2010). Improving your data transformations : Applying the Box-Cox transformation. 15(12):1–9.

Pickering, J. W., Frampton, C. M., and Endre, Z. H. (2009). Evaluation of trial outcomes in acute kidney injury by creatinine modeling. *Clinical journal of the American Society of Nephrology : CJASN*, 4(11):1705–15.

Rigby, B. and Stasinopoulos, M. (2009). *A flexible regression approach using GAMLSS in R* .

Rigby, B., Stasinopoulos, M., Heller, G., and Voudouris, V. (2014). *The Distribution Toolbox of GAMLSS*.

Schimek, M. G. (2000). *Smoothing and Regression - Approaches, Computation and Application*. wiley Series in Probability and Statistics.

Siew, E. D., Matheny, M. E., Ikizler, T. A., Lewis, J. B., Miller, R. A., Waitman, L. R., Go, A. S., Parikh, C. R., and Peterson, J. F. (2010). Commonly used surrogates for baseline renal function affect the classification and prognosis of acute kidney injury. *International Society of Nephrology*, 77:536–542.

Silva, A. L. T. P. (2006). *Modelos Aditivos Generalizados em Análise de Sobrevivência*. PhD thesis, Faculdade de Ciências da Universidade de Lisboa.

Silva, M. J. J. (2012). *Modelo de prognóstico do tempo necessário ao controlo da dor oncológica*. Master thesis, Faculdade de Ciências da Universidade de Lisboa.

Solomon, R. and Segal, A. (2008). Defining acute kidney injury: what is the most appropriate metric? *Nature clinical practice. Nephrology*, 4(4):208–15.

Soto, K., Coelho, S., Rodrigues, B., Martins, H., Frade, F., Lopes, S., Cunha, L., Papoila, A. L., and Devarajan, P. (2010). Cystatin C as a Marker of Acute Kidney Injury in the Emergency Department. *Clinical journal of the American Society of Nephrology : CJASN*, 5(10):1745–54.

Soto, K., Papoila, A. L., Coelho, S., Bennett, M., Ma, Q., Rodrigues, B., Fidalgo, P., Frade, F., and Devarajan, P. (2013). Plasma NGAL for the diagnosis of AKI in patients admitted from the emergency department setting. *Clinical journal of the American Society of Nephrology : CJASN*, 8(12):2053–63.

Spedicato, G. A., Clemente, G. P., and Schewe, F. (2014). The Use of GAMLSS in Assessing the Distribution of Unpaid Claims Reserves. *Casualty Actuarial Society E-Forum*.

Stasinopoulos, M. and Rigby, B. (2014). The GAMLSS packages in R.

Stasinopoulos, M., Rigby, B., Vlasios, V., Heler, G., and Fernanda, B. (2015). Flexible Regression and Smoothing The GAMLSS packages in R.

Turkman, M. A. A. and Silva, G. L. (2000). *Modelos Lineares Generalizados - da teoria à prática*. Edições SPE.

Waikar, S. S., Betensky, R. a., and Bonventre, J. V. (2009). Creatinine as the gold standard for kidney injury biomarker studies? *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, 24(11):3263–5.

Waikar, S. S. and Bonventre, J. V. (2009). Creatinine kinetics and the definition of acute kidney injury. *Journal of the American Society of Nephrology : JASN*, 20(3):672–9.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. Taylor & Francis Group.

Závada, J., Hoste, E., Cartin-Ceba, R., Calzavacca, P., Gajic, O., Clermont, G., Bellomo, R., and Kellum, J. a. (2010). A comparison of three methods to estimate baseline creatinine for RIFLE classification. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, 25(12):1–8.

Apêndice

Todos os resultados apresentados neste trabalho resultaram de análises efetuadas através do *software* R. Assim, todo o código construído e todas as funções e *packages* utilizados neste trabalho encontram-se apresentados nesta secção.

```
#############################################################################
#############################################################################
######################     Abertura dos Dados     ######################
#############################################################################
#############################################################################

require(foreign)
dados<-read.spss('/Users/Inês/Desktop/base_corrigida_kdigo.sav',
use.value.labels=TRUE,to.data.frame=TRUE)

#############################################################################
#############################################################################
############################     Packages     ###########################
#############################################################################
#############################################################################

library(gamlss)
library(AID)
library(gam)

#############################################################################
#############################################################################
############################     Variáveis     ###########################
#############################################################################
#############################################################################



########################## Variáveis Independentes

####### Sexo

sexo<-dados$SEXfemale2male1
sexo<-factor(sexo, levels = c(1,2), labels = c("male", "female"))
summary(sexo)
```

```
####### Idade

idade<-dados$AGE
summary(idade)
sd(idade)

####### Raça

raca<-dados$RaceNonblack1black2
raca<-factor(raca, levels = c(1,2), labels = c("nonBlack", "Black"))
summary(raca)

####### Creatinina no tempo 0

scr_t0<-dados$Scr_T0
summary(scr_t0)
sd(scr_t0)

########################## Variável Dependente - Creatinina sérica basal

scrb<-dados$BASELINEScr
summary(scrb)
sd(scrb)

### Transformação logarítmica da variável dependente  - log(SCrB)

log_scrb<-log(scrb)
summary(log_scrb)
sd(log_scrb)

##############################################################################
##############################################################################
##################     Data Frame dos dados     ##########################
##############################################################################
##############################################################################

data_frame_dados<-data.frame(scrb,idade,sexo,raca,scr_t0,log_scrb)
dados_finais<-data_frame_dados
```

```
#########################################################################
#########################################################################
#######    Teste de Normalidade através do Kolmogorov-Smirnov    ######
#########################################################################
#########################################################################

########################## Creatinina sérica basal

ks.test(scrb, "pnorm",mean(scrb),sd(scrb))

########################## LOG Creatinina sérica basal

ks.test(log_scrb, "pnorm",mean(log_scrb),sd(log_scrb))

########################## Sexo, Creatinina sérica basal
par(mfrow=c(2,2))
subset_feminino<-subset(dados_finais,sexo=='female')
ks.test(subset_feminino[,1], "pnorm",mean(subset_feminino[,1]),
sd(subset_feminino[,1]))

summary(subset_feminino$scrb)
sd(subset_feminino$scrb)

qqnorm(subset_feminino[,1], main='QQ-plot SCr Basal - Sexo Feminino')
abline(0,1,col='red')

subset_masculino<-subset(dados_finais,sexo=='male')
ks.test(subset_masculino[,1], "pnorm",mean(subset_masculino[,1]),
sd(subset_masculino[,1]))

summary(subset_masculino$scrb)
sd(subset_masculino$scrb)


qqnorm(subset_masculino[,1], main='QQ-plot SCr Basal - Sexo Masculino')
abline(0,1,col='red')
```

```
########################## Raça, Creatinina sérica basal

subset_nonBLack<-subset(dados_finais,raca=='nonBlack')
ks.test(subset_nonBLack[,1], "pnorm",mean(subset_nonBLack[,1]),
sd(subset_nonBLack[,1]))

summary(subset_nonBLack$scrb)
sd(subset_nonBLack$scrb)

qqnorm(subset_nonBLack[,1], main='QQ-plot SCr Basal - Raça não Negra')
abline(0,1,col='red')

subset_Black<-subset(dados_finais,raca=='Black')
ks.test(subset_Black[,1], "pnorm",mean(subset_Black[,1]),
sd(subset_Black[,1]))

summary(subset_Black$scrb)
sd(subset_Black$scrb)


qqnorm(subset_Black[,1], main='QQ-plot SCr Basal - Raça Negra')
abline(0,1,col='red')



##############################################################################
##############################################################################
##########################    QQ NORM PLOT   ###############################
##############################################################################
##############################################################################

par(mfrow=c(1,1))

########################## Creatinina sérica basal

qqnorm(scrb,main='QQ-plot SCr Basal')
abline(0,1,col='red')

hist(scrb, main='Histograma da SCrB', ylab='Frequências', xlab='SCrB',
col=c('blue'))
```

```
########################## Raça, Creatinina sérica basal

subset_nonBLack<-subset(dados_finais,raca=='nonBlack')
ks.test(subset_nonBLack[,1], "pnorm",mean(subset_nonBLack[,1]),
sd(subset_nonBLack[,1]))

summary(subset_nonBLack$scrb)
sd(subset_nonBLack$scrb)

qqnorm(subset_nonBLack[,1], main='QQ-plot SCr Basal - Raça não Negra')
abline(0,1,col='red')

subset_Black<-subset(dados_finais,raca=='Black')
ks.test(subset_Black[,1], "pnorm",mean(subset_Black[,1]),
sd(subset_Black[,1]))

summary(subset_Black$scrb)
sd(subset_Black$scrb)


qqnorm(subset_Black[,1], main='QQ-plot SCr Basal - Raça Negra')
abline(0,1,col='red')



##############################################################################
##############################################################################
##########################    QQ NORM PLOT    ################################
##############################################################################
##############################################################################

par(mfrow=c(1,1))

########################## Creatinina sérica basal

qqnorm(scrb,main='QQ-plot SCr Basal')
abline(0,1,col='red')

hist(scrb, main='Histograma da SCrB', ylab='Frequências', xlab='SCrB',
col=c('blue'))
```

```
########################## Logarítmo Creatinina sérica basal

qqnorm(log_scrb,main='QQ-plot Log_SCr Basal')
abline(0,1,col='red')

hist(log_scrb, main='Histograma da LOG_SCrB', ylab='Frequências',
xlab='LOG_SCrB', col=c('darkorange1'))


###############################################################################
###############################################################################
###   Testar se existem diferenças dos valores médios de SCrB entre SEXOS   ###
###############################################################################
###############################################################################

scrb_sexo_masculino<-scrb[sexo=='male']
scrb_sexo_feminino<-scrb[sexo=='female']
par(mfrow=c(1,2))


teste_t_sexo<-t.test(scrb_sexo_masculino,scrb_sexo_feminino,var.equal=T)
teste_t_sexo


###############################################################################
###############################################################################
###   Testar se existem diferenças dos valores médios de SCrB entre RAÇAS   ###
###############################################################################
###############################################################################

scrb_raca_nonBlack<-scrb[raca=='nonBlack']
scrb_raca_Black<-scrb[raca=='Black']
par(mfrow=c(1,2))

teste_t_raca<-t.test(scrb_raca_nonBlack,scrb_raca_Black,var.equal=F)
teste_t_raca

par(mfrow=c(1,1))
```

```
##############################################################################
##############################################################################
#############   MODELOS Lineares Generalizados - GLM   ################
##############################################################################
##############################################################################

glm_model<-glm(scrb~scr_t0+sexo+idade+raca,
family=gaussian, data=dados_finais)


########################### RESIDUOS do GLM

residuos_glm<-residuals(glm_model)

summary(residuos_glm)


qqnorm(residuos_glm, main='QQ-plot Resíduos - MLG')
abline(0,1,col='red')


##############################################################################
##############################################################################
#############   MODELOS Lineares Generalizados - Log(GLM)   ################
##############################################################################
##############################################################################

glm_model_log_scrb<-glm(log_scrb~scr_t0+sexo+idade+raca,
family=gaussian, data=dados_finais)

########################### RESIDUOS do log(GLM)

residuos_glm_log<-residuals(glm_model_log_scrb)

qqnorm(residuos_glm_log, main='QQ plot Resíduos - MLG_log(SCrB)')
abline(0,1,col='red')
```

```
##############################################################################
##############################################################################
#####################   Transformação Box-Cox   ##############################
##############################################################################
##############################################################################

boxcoxnc(scrb,method="all")

##############################################################################
##############################################################################
#############   MODELOS Aditivos Generalizados - GAM   #######################
##############################################################################
##############################################################################

##### MODELOS UNIVARIADOS

####   IDADE

gam_model_idade<-gam(scrb~s(idade), family=gaussian)

####   CREATININA TEMPO 0

gam_model_scr_t0<-gam(scrb~s(scr_t0), family=gaussian)


par(mfrow=c(1,2))

plot (gam_model_idade, se=TRUE,main='Modelo aditivo Univariado
 - SCrB~s(Idade)')
plot (gam_model_scr_t0, se=T,main='Modelo aditivo Univariado
- SCrB~s(SCr_t0)')


#### MODELO MULTIVARIADO

modelo_gam<-gam(scrb~s(scr_t0)+ sexo + idade + raca,
family=gaussian, data= dados_finais, na.action=na.omit)
```

```
########################## RESIDUOS do GAM

residuos_gam<-residuals(modelo_gam)

par(mfrow=c(1,1))

qqnorm(residuos_gam, main='QQ-plot Resíduos - MAG')
abline(0,1,col='red')


#########################################################################
#########################################################################
######################    MODELO - GAMLSS    ############################
#########################################################################
#########################################################################

#### DISTRIBUICAO AUTOMATICO

fitting_dist<-fitDist(scrb,type=c('realplus'),
extra=c('BCPE','BCCG','BCT','GB2', 'WEI','WEI2'))

fitting_dist$fits
fitting_dist$failed

histDist(scrb,family = "GB2",nbins=30,density=TRUE,ylim=c(0,2.5))

#########################################################################
#########################################################################
###############     MODELOS UNIVARIADOS - GAMLSS     ####################
#########################################################################
#########################################################################

###################### IDADE

modelo_gamlss_idade<-gamlss(scrb~idade,family="GB2",n.cyc=220)
```

```
###################### CREATININA SERICA TEMPO 0

modelo_gamlss_scr_t0<-gamlss(scrb~cs(scr_t0),family="GB2",n.cyc=220)
modelo_gamlss_scr_t0<-gamlss(scrb~pvc(scr_t0),family="GB2",n.cyc=740)
modelo_gamlss_scr_t0<-gamlss(scrb~cy(scr_t0),family="GB2",n.cyc=1000)
modelo_gamlss_scr_t0<-gamlss(scrb~pb(scr_t0),family="GB2",n.cyc=740)
modelo_gamlss_scr_t0<-gamlss(scrb~ps(scr_t0),family="GB2",n.cyc=170)
modelo_gamlss_scr_t0<-gamlss(scrb~ri(scr_t0),family="GB2",n.cyc=200)
modelo_gamlss_scr_t0<-gamlss(scrb~fp(scr_t0),family="GB2",n.cyc=340)
modelo_gamlss_scr_t0<-gamlss(scrb~scs(scr_t0),family="GB2",n.cyc=900)
modelo_gamlss_scr_t0<-gamlss(scrb~lo(~scr_t0),family="GB2",n.cyc=100000)

##########################################################################
##########################################################################
##############     MODELO MULTIVARIADO - GAMLSS      ###################
##########################################################################
##########################################################################

######################### SELEÇÃO DAS VARIÁVEIS PARA O MODELO

############ stepGAIC

modelo_gamlss_completo_todos_parametros<-gamlss(scrb~scs(scr_t0)+sexo+
idade+raca,family="GB2",n.cyc=10000,
sigma.formula =~scs(scr_t0)+sexo+idade+raca,
nu.formula =~scs(scr_t0)+sexo+idade+raca,
tau.formula=~scs(scr_t0)+sexo+idade+raca)

step_completo_scope_todos_parametros<-stepGAIC(
modelo_gamlss_completo_todos_parametros,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca))

step_completo_scope_sigma_todos_parametros<-stepGAIC(
modelo_gamlss_completo_todos_parametros,scope=list(
lower=~1, upper=~scs(scr_t0)+
sexo+idade+raca),what='sigma')

step_completo_scope_nu_todos_parametros<-stepGAIC(
modelo_gamlss_completo_todos_parametros,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca),
what='nu')

step_completo_scope_tau_todos_parametros<-stepGAIC(
modelo_gamlss_completo_todos_parametros,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca),
what='tau')
```

```
############# stepGAICAll.A

stepGAIC.A_completo_scope_todos_parametros<-stepGAICAll.A(
modelo_gamlss_completo_todos_parametros,scope=list(lower=~1,
upper=~scs(scr_t0)+sexo+idade+raca),steps=1000000)

############# stepGAICAll.B

stepGAIC.B_completo_scope_todos_parametros<-stepGAICAll.B(
modelo_gamlss_completo_todos_parametros,scope=list(lower=~1,
upper=~scs(scr_t0)+sexo+idade+raca),steps=1000000)

###########################################################################
###########################################################################
###################     MODELO MULTIVARIADO - GAMLSS      #################
###########################################################################
###########################################################################
###################      Modelação pelo stepGAIC()       ##################
###########################################################################
###########################################################################

modelo_gamlss_todos<-gamlss(scrb~idade+scs(scr_t0)+raca+sexo,
sigma.fo=~idade+scs(scr_t0)+sexo,nu.fo=~scs(scr_t0),
tau.fo=~idade+raca+sexo,family="GB2",n.cyc=6000,data=dados_finais)

par(mfrow=c(1,1))

summary(modelo_gamlss_todos)
plot(modelo_gamlss_todos)


########################## Histograma dos valores estimados, através do GAMLSS

estimados_gamlss_todos<-fitted(modelo_gamlss_todos)

par(mfrow=c(1,2))

hist(estimados_gamlss_todos, main='Valores estimados da SCrB - GAMLSS',
ylab='Frequências', xlab='Valores estimados', col=c('orange1'),
xlim=c(0,3),ylim=c(0,350))

hist(scrb, main='Valores observados da SCrB', ylab='Frequências',
xlab='Valores observados', col=c('blue'),xlim=c(0,3),ylim=c(0,350))
```

```
########################## Plot
par(mfrow=c(1,1))

plot(scrb,estimados_gamlss_todos,
main='Gráfico dos valores estimados e observados da creatinina sérica basal',
xlab='Valores observados da SCrB',ylab='Valores estimados da SCrB pelo GAMLSS',
xlim=c(0,3),ylim=c(0,3))

abline(0,1,col='red')

########################## RESIDUOS do GAMLSS
residuos_gamlss_todos<-residuals(modelo_gamlss_todos)

par(mfrow=c(1,2))

boxplot(residuos_gamlss_todos,col='yellow',main='Boxplot Resíduos - GAMLSS',
xlab='Resíduos')
abline(h=0,col='red')

qqnorm(residuos_gamlss, main='QQ-plot Resíduos - GAMLSS')
abline(0,1,col='red')

par(mfrow=c(1,1))

wp(modelo_gamlss_todos,main='Worm plot dos Resíduos')

##############################################################################
##############################################################################
####################     MODELO MULTIVARIADO - GAMLSS      ###################
##############################################################################
##############################################################################
#######################      Avaliação das Interações       ##################
##############################################################################
##############################################################################

modelo_gamlss_todos_inter1<-gamlss(scrb~idade*sexo+scs(scr_t0)+raca,
sigma.fo=~idade*sexo+scs(scr_t0),nu.fo=~scs(scr_t0),
tau.fo=~idade*sexo+raca,family="GB2",n.cyc=6000,data=dados_finais)

summary(modelo_gamlss_todos_inter1)


modelo_gamlss_todos_inter2<-gamlss(scrb~idade*raca+scs(scr_t0)+sexo,
sigma.fo=~idade+scs(scr_t0)+sexo,nu.fo=~scs(scr_t0),
tau.fo=~idade*raca+sexo,family="GB2",n.cyc=6000,data=dados_finais)

summary(modelo_gamlss_todos_inter2)
```

```
modelo_gamlss_todos_inter3<-gamlss(scrb~idade+scs(scr_t0)+raca*sexo,
sigma.fo=~idade+scs(scr_t0)+sexo,nu.fo=~scs(scr_t0),
tau.fo=~idade+raca*sexo,family="GB2",n.cyc=6000,data=dados_finais)

summary(modelo_gamlss_todos_inter3)

modelo_gamlss_todos_inter4<-gamlss(scrb~scs(scr_t0)+idade*raca*sexo,
sigma.fo=~idade*sexo+scs(scr_t0),nu.fo=~scs(scr_t0),
tau.fo=~idade*raca*sexo,family="GB2",n.cyc=6000,data=dados_finais)

summary(modelo_gamlss_todos_inter4)

par(mfrow=c(1,1))

##############################################################################
##############################################################################
###################     MODELO MULTIVARIADO - GAMLSS      ##################
##############################################################################
##############################################################################
###############      Modelação pelo stepGAICAll.B()       ##################
##############################################################################
##############################################################################

modelo_gamlss_todos_B<-gamlss(scrb~scs(scr_t0)+sexo,
sigma.fo=~scs(scr_t0)+sexo,nu.fo=~scs(scr_t0)+sexo,
tau.fo=~scs(scr_t0)+sexo,family="GB2",n.cyc=6000,data=dados_finais)

summary(modelo_gamlss_todos_B)
plot(modelo_gamlss_todos_B)

########################### Histograma dos valores estimados, através do GAMLSS

estimados_gamlss_todos_B<-fitted(modelo_gamlss_todos_B)

par(mfrow=c(1,2))

hist(estimados_gamlss_todos_B, main='Valores estimados da SCrB - GAMLSS',
ylab='Frequências', xlab='Valores estimados', col=c('orange1'))

hist(scrb, main='Valores observados da SCrB', ylab='Frequências',
xlab='Valores observados', col=c('blue'),xlim=c(0,3),ylim=c(0,350))

########################### Plot
par(mfrow=c(1,1))

plot(scrb,estimados_gamlss_todos_B,
main='Gráfico dos valores estimados e observados da creatinina sérica basal',
xlab='Valores observados da SCrB',ylab='Valores estimados da SCrB pelo GAMLSS')
```

```
abline(0,1,col='red')

########################## RESIDUOS do GAMLSS

residuos_gamlss_todos_B<-residuals(modelo_gamlss_todos_B)

par(mfrow=c(1,2))

boxplot(residuos_gamlss_todos_B,col='yellow',main='Boxplot Resíduos - GAMLSS',
xlab='Resíduos')
abline(h=0,col='red')

qqnorm(residuos_gamlss_todos_B, main='QQ-plot Resíduos - GAMLSS')
abline(0,1,col='red')

par(mfrow=c(1,1))

wp(modelo_gamlss_todos_B,main='Worm plot dos Resíduos')

##############################################################################
##############################################################################
###################     MODELO MULTIVARIADO - GAMLSS      ####################
##############################################################################
##############################################################################
###################      Modelação pelo stepGAIC()       #####################
##############################################################################
##############################################################################
#############################    sem outliers   ##############################
##############################################################################

###### Identificação dos outliers GAMLSS

valores_outliers<- boxplot.stats(residuos_gamlss_todos)$out
residuos_gamlss_todos<-round(residuos_gamlss_todos,6)

which(residuos_gamlss_todos==-2.851965) #30
which(residuos_gamlss_todos==2.729684) #298
which(residuos_gamlss_todos==3.082837) #299
which(residuos_gamlss_todos==-2.751153) #419
which(residuos_gamlss_todos==2.656329) #526
which(residuos_gamlss_todos==2.916072) #564

dados_finais_sem_out_gamlss<-dados_finais[-c(30, 298, 299,419,526,564), ]
scrb_out<- scrb[-c(30, 298, 299,419,526,564)]


fitting_dist_out<-fitDist(dados_finais_sem_out_gamlss[,1],
type=c('realplus'),extra=c('BCPE','BCCG','BCT','GB2', 'WEI','WEI2'))
```

```
fitting_dist_out$fits
fitting_dist_out$failed

histDist(dados_finais_sem_out_gamlss$scrb,family = "GB2",nbins=30,
density=TRUE,ylim=c(0,2.5))


################################################
###### Seleção das variáveis pelo stepGAIC()
################################################

modelo_gamlss_completo_todos_parametros_out<-gamlss(scrb~scs(scr_t0)+sexo+
idade+raca,family="GB2",n.cyc=10000,
sigma.formula =~scs(scr_t0)+sexo+idade+raca,
nu.formula =~scs(scr_t0)+sexo+idade+raca,
tau.formula=~scs(scr_t0)+sexo+idade+raca, data=dados_finais_sem_out_gamlss)

step_completo_scope_todos_parametros_out<-stepGAIC(
modelo_gamlss_completo_todos_parametros_out,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca),
data=dados_finais_sem_out_gamlss)

step_completo_scope_sigma_todos_parametros_out<-stepGAIC(
modelo_gamlss_completo_todos_parametros_out,scope=list(
lower=~1, upper=~scs(scr_t0)+
sexo+idade+raca),what='sigma',
data=dados_finais_sem_out_gamlss)


step_completo_scope_nu_todos_parametros_out<-stepGAIC(
modelo_gamlss_completo_todos_parametros_out,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca),
what='nu',data=dados_finais_sem_out_gamlss)

step_completo_scope_tau_todos_parametros_out<-stepGAIC(
modelo_gamlss_completo_todos_parametros_out,scope=list(
lower=~1, upper=~scs(scr_t0)+sexo+idade+raca),
what='tau',data=dados_finais_sem_out_gamlss)


################################################
########### Modelo GAMLSS sem valores extremos
################################################

modelo_gamlss_todos_out<-gamlss(scrb~idade+scs(scr_t0)+raca,
sigma.fo=~idade+scs(scr_t0)+sexo,nu.fo=~scs(scr_t0),
tau.fo=~idade+raca+sexo,family="GB2",n.cyc=600,
data=dados_finais_sem_out_gamlss)
```

```
summary(modelo_gamlss_todos_out)
plot(modelo_gamlss_todos_out)


########################## Histograma dos valores estimados, através do GAMLSS

estimados_gamlss_todos_out<-fitted(modelo_gamlss_todos_out)

par(mfrow=c(1,2))

hist(estimados_gamlss_todos_out, main='Valores estimados da SCrB - GAMLSS',
ylab='Frequências', xlab='Valores estimados', col=c('orange1'),
xlim=c(0,3),ylim=c(0,350))

hist(scrb_out, main='Valores observados da SCrB', ylab='Frequências',
xlab='Valores observados', col=c('blue'),xlim=c(0,3),ylim=c(0,350))

########################## Plot
par(mfrow=c(1,1))

plot(scrb_out,estimados_gamlss_todos_out,
main='Gráfico dos valores estimados e observados da creatinina sérica basal',
xlab='Valores observados da SCrB',ylab='Valores estimados da SCrB pelo GAMLSS',
xlim=c(0,3),ylim=c(0,3))

abline(0,1,col='red')

########################## RESIDUOS do GAMLSS
residuos_gamlss_todos_out<-residuals(modelo_gamlss_todos_out)

par(mfrow=c(1,2))

boxplot(residuos_gamlss_todos_out,col='yellow',main='Boxplot Resíduos - GAMLSS',
xlab='Resíduos')
abline(h=0,col='red')

qqnorm(residuos_gamlss_todos_out, main='QQ-plot Resíduos - GAMLSS')
abline(0,1,col='red')

par(mfrow=c(1,1))

wp(modelo_gamlss_todos_out,main='Worm plot dos Resíduos',ylim.all=0.6)
```